UNIVERSITÀ DEGLI STUDI DI PADOVA

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione finale

Un coefficiente di determinazione per modelli lineari generalizzati

Relatore: Prof. Euloge Clovis Kenne Pagui

Dipartimento di Scienze Statistiche

Laureando: Riccardo Guderzo
Matricola: 1147857

2018/2019

Indice

Introduzione

La statistica ha l'importante compito di ottenere informazione attraverso l'esame di dati empirici e la modellazione matematica della loro variabilità. A tal fine vengono adottate diverse procedure con l'obiettivo di raccogliere, analizzare, interpretare, sintetizzare e presentare i dati. Durante queste fasi, la controversia di maggior rilievo è data dalla selezione del modello: non esiste infatti il modello *perfetto* che riesce a spiegare il fenomeno senza alcun margine d'errore. Obiettivo della modellazione statistica sarà pertanto quello di cercare il modello *migliore* tra quelli selezionabili.
Sotto assunzione di gaussianità, tra gli indici utilizzati durante la fase di creazione del modello, il coefficiente di determinazione (anche detto R^2) è senz'altro una delle misure più popolari per valutare la bontà di adattamento di un modello lineare grazie alla sua semplicità e intuitività, come affermato da Draper e Smith (1998). Le estensioni del modello lineare, però, sono molto frequenti nell'analisi statistica e, se per i modelli lineari la bontà di adattamento è facilmente misurabile tramite tale indice, così non è per i modelli lineari generalizzati in quanto non viene rispettata l'assunzione di gaussianità dei termini di errore. Sono quindi stati proposti numerosi indici ma ad oggi non è ancora emerso nessun consenso su quale sia il miglior approccio in questo contesto.
Scopo di tale lavoro sarà quello di confrontare, tramite simulazione e applicazione a dei *data-set* reali, diversi coefficienti di determinazione per modelli lineari generalizzati. In particolare verrà spiegata dettagliatamente la teoria alla base di questi indici e verranno testate le loro prestazioni nella spiegazione della bontà di adattamento del modello usando il *software R*. Particolare attenzione sarà posta al coefficiente di determinazione proposto da Zhang (2017), cercando di trovarne pregi, difetti e miglioramenti rispetto agli R^2 proposti precedentemente.
Per conseguire l'obiettivo stabilito si daranno innanzitutto dei concetti generali sull'inferenza statistica al fine di stabilire la notazione a cui si farà riferimento e per delineare le relazioni che queste hanno con l'argomento di interesse. Il Capitolo 1 coprirà quindi tutta la teoria alla base dell'argomento proposto introducendo la verosimiglianza e tutta la teoria che concerne l'adattamento di modelli (lineari e non). Il Capitolo 2 avrà l'obiettivo di chiarire la teoria alla base dei coefficienti di determinazione per modelli lineari generalizzati. Nel Capitolo 3 verranno testati, tramite simulazione, tali coefficienti per confrontare e valutare la loro efficacia in termini di spiegazione della bontà di

adattamento del modello. Nel Capitolo 4 verranno verificati gli R^2 su *data-set* reali. Verranno infine tratte le dovute conclusioni sull'efficacia del coefficiente di determinazione di Zhang nella spiegazione della bontà di adattamento del modello.

Capitolo 1

Modelli parametrici e verosimiglianza

1.1 Introduzione

In questo capitolo si vuole fornire una rapida panoramica dei concetti che stanno alla base dell'inferenza statistica basata sulla verosimiglianza. Successivamente si discuterà dei modelli statistici concentrandosi particolarmente sui modelli lineari generalizzati. Le informazioni qui fornite sono basate su Pace e Salvan (2001), su Grigoletto, Pauli e Ventura (2017) e su Dunn e Smyth (2018).

1.2 Inferenza di verosimiglianza

1.2.1 Specificazione del modello

Un modello statistico deve descrivere nel migliore dei modi il fenomeno di diretto interesse di una popolazione basandosi su un campione limitato di osservazioni $(y_1, y_2, ..., y_n)$ realizzazione di una variabile casuale $Y = (Y_1, Y_2, ..., Y_n)$ a componenti indipendenti. Assumendo che Y sia generata da un modello con legge di probabilità $f_0(y)$ e spazio campionario $\mathcal{Y}$, l'obiettivo sarà quello di ottenere delle conclusioni su $f_0(y)$ basandosi sulla sola informazione data dal campione. Sarà necessario definire, prima di tutto, il modello statistico $\mathcal{F}$, ovvero una classe di distribuzioni a cui viene limitato lo studio. Se il modello è correttamente specificato, allora $f_0(y)$ appartiene a $\mathcal{F}$. Si può quindi definire $\mathcal{F}$ come un modello parametrico, ovvero

$$\mathcal{F} = \{f(y;\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\} \tag{1.1}$$

con $d \in \mathbb{N}^+$ e Θ spazio parametrico, cioè lo spazio contenente tutti i possibili valori di θ parametro d-dimensionale. Si assume che θ sia identificabile, ossia che la corrispondenza tra Θ e $\mathcal{F}$ sia biunivoca.

1.2.2 Funzione di verosimiglianza

L'inferenza è uno dei punti cruciali della statistica. Fare inferenza significa che, una volta definito il modello statistico $\mathcal{F}$, sarà d'interesse valutare quali sono gli elementi appartenenti a $\mathcal{F}$ che sono più verosimili alla luce del campione osservato.
A tal fine, assunto che il modello $\mathcal{F}$ con densità $f(y;\theta)$ sia correttamente specificato, si definisce la funzione di verosimiglianza (*likelihood function*) $L : \Theta \to \mathbb{R}^+$ come

$$L(\theta) = L(\theta; y) = c(y)f(y;\theta), \tag{1.2}$$

con $c(y)$ funzione dei dati e indipendente dal parametro θ. Rispetto al modello $\mathcal{F}$, la funzione di verosimiglianza è dunque una classe di funzioni equivalenti che differiscono tra loro per una sola componente $c(y)$ costante. Il modello statistico per i dati $(y_1, y_2, ..., y_n)$ assume che essi siano n realizzazioni indipendenti, pertanto la funzione di verosimiglianza è il semplice prodotto delle singole densità, si può cioè esprimere $L(\theta)$ come

$$L(\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i, \theta),$$

con $f_{Y_i}(y_i, \theta)$ funzione di densità della variabile Y_i generatrice della i-esima osservazione, y_i, del campione. Più in generale, la verosimiglianza permette di ottenere informazione su un parametro ignoto partendo da più esperimenti indipendenti. La funzione di verosimiglianza complessiva non è altro che il prodotto delle funzioni di verosimiglianza nei singoli esperimenti.
Al fine di semplificare l'esecuzione pratica dei calcoli e rendere più semplice la rappresentazione dei risultati teorici si fa spesso ricorso al logaritmo naturale della funzione di verosimiglianza. Essa viene definita funzione di log-verosimiglianza (*log-likelihood function*) definita come

$$l(\theta) = l(\theta; y) = logL(\theta; y), \tag{1.3}$$

pertanto sotto campionamento casuale semplice (indipendenza dei dati) la log- verosimiglianza assumerà la forma

$$l(\theta) = \sum_{i=1}^{n} l(\theta; y_i).$$

1.3 Stimatore di massima verosimiglianza

Un valore $\hat{\theta} \in \Theta$ per cui $L(\hat{\theta}) > L(\theta)$ per ogni $\theta \in \Theta$ è detto stima di massima verosimiglianza (SMV) del vero e ignoto parametro θ_0. Essendo il logaritmo una funzione monotona, $\hat{\theta}$ può essere determinato anche a partire dalla funzione di log-verosimiglianza. Se $\hat{\theta} = \hat{\theta}(y)$ esiste ed è unico con probabilità uno, la variabile casuale $\hat{\theta} = \hat{\theta}(Y)$ è detta stimatore di massima verosimiglianza. Per trovarne una stima, dunque, basta sostituire le osservazioni $(y_1, y_2, ..., y_n)$ con il vettore casuale $(Y_1, Y_2, ..., Y_n)$.

Nella maggioranza dei casi i modelli considerati nelle applicazioni seguono delle condizioni necessarie per l'applicazione dei metodi che si andranno a discutere. Sotto tali vincoli il modello è detto con verosimiglianza regolare. In particolare se:

- Θ è sottoinsieme aperto di $\mathbb{R}^d$;
- $l(\theta)$ è differenziabile almeno tre volte in θ con derivate parziali continue;
- il supporto della densità del modello è indipendente dal parametro θ;

allora il modello statistico parametrico scelto è verosimilmente regolare.

Le condizioni di regolarità permettono di ottenere la stima di massima verosimiglianza come punto di stazionarietà della log-verosimiglianza. In particolare, dato il parametro $\theta = (\theta_1, \theta_2, ..., \theta_d)$, il vettore delle derivate parziali corrispondente all'insieme d-dimensionale delle equazioni di verosimiglianza

$$l_*(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, ..., \frac{\partial l(\theta)}{\partial \theta_d} \right)^\top, \tag{1.4}$$

è detto funzione *score* o funzione punteggio. Il generico elemento di $l_*(\theta)$ è $l_r(\theta) = \partial l(\theta)/\partial \theta_r$, $r = 1, ..., d$.

La matrice $d \times d$ delle derivate parziali seconde di $l(\theta)$ cambiate di segno,

$$j(\theta) = -l_{**}(\theta) = -\frac{\partial^2}{\partial \theta \partial \theta^\top} l(\theta), \tag{1.5}$$

è detta matrice di informazione osservata. Il generico elemento di $j(\theta)$ è $j_{rs} = -\partial l(\theta)/\partial \theta_r \partial \theta_s$, $r, s = 1, ..., d$. Questa matrice descrive la curvatura locale della probabilità attorno al suo massimo. Pertanto, quanto più grande è $j(\hat{\theta})$, tanto più la verosimiglianza è concentrata attorno a $\hat{\theta}$. Quindi $j(\hat{\theta})$ è una misura dell'informazione ottenuta dai dati sull'incognito parametro θ.

Il valore atteso sotto θ della matrice di informazione osservata

$$i(\theta) = E_\theta[j(\theta)] \tag{1.6}$$

è detto matrice di informazione attesa ed è un'ulteriore quantità notevole di verosimiglianza.

1.4 Verosimiglianza e proprietà campionarie

In modelli con verosimiglianza regolare, in particolare per modelli in cui il supporto di Y non dipende da θ, valgono i seguenti risultati esatti:

Prima identità di Bartlett

$$E_\theta(l_*(\theta; Y)) = 0 \quad \forall \theta \in \Theta. \tag{1.7}$$

Ciò significa che se viene fatta la media rispetto al parametro θ all'interno dell'espressione, allora l_* ha media nulla, ciò vale per ogni θ fissato. Dunque, in generale:

$$E_{\theta 1}(l_*(\theta_2; Y)) \neq 0.$$

Seconda identità di Bartlett

$$Var_\theta(l_*(\theta; Y)) = E_\theta(l_*(\theta; Y)(l_*(\theta; Y))^\top) = i(\theta) \quad \forall \theta \in \Theta. \tag{1.8}$$

La (1.8) è anche detta identità dell'informazione.
Si vuole rendere noto anche che con $y = (y_1, ..., y_n)$ sotto condizioni di regolarità, sono disponibili utili approssimazioni asintotiche per lo stimatore di massima verosimiglianza definito nel paragrafo precedente. Non si vuole comunque approfondire ulteriormente l'argomento in quanto potrebbe risultare fuorviante per lo scopo di questo lavoro.

1.5 Il modello di regressione lineare

I modelli di regressione lineare giocano un ruolo chiave nello studio delle relazioni tra variabili e costituiscono lo schema su cui poggiano metodi più complessi. Essi permettono di studiare la relazione tra una variabile risposta Y e una o più variabili esplicative $(x_1, ..., x_p)$, con $p \geqslant 1$.

1.5.1 Assunzioni

Il modello di regressione lineare semplice poggia sulle seguenti assunzioni:

1. $Y = X\beta + \epsilon = \eta + \epsilon$, con

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & ... & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & ... & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \eta = X\beta.$$

2. X è una matrice di costanti $n \times p$, $p < n$, con righe $x_1, ..., x_n$ e rango pieno p.

3. $\epsilon \sim N_n(0, \sigma^2 I_n)$, con $\sigma^2 > 0$.

La notazione $N_n(\mu, \Sigma)$ indica una distribuzione normale n-dimensionale con vettore delle medie μ e matrice di covarianza Σ.

1.5.2 Inferenza sui parametri

Nei modelli di regressione lineare normale la funzione di log-verosimiglianza per i parametri (β, σ^2) è definita sullo spazio parametrico $\mathbb{R}^p \times (0, +\infty)$ da

$$\begin{aligned} l(\beta, \sigma^2; y) &= -\frac{n}{2} log\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 \\ &= -\frac{n}{2} log\sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \\ &= -\frac{n}{2} log\sigma^2 - \frac{1}{2\sigma^2} \|y - X\beta\|^2, \end{aligned}$$

dove, dato un vettore $u \in \mathbb{R}^n$, $\|u\|^2 = u^\top u$ è il quadrato della norma di u.
La stima di massima verosimiglianza per il parametro p-dimensionale β sarà pari a

$$\hat{\beta} = (X^\top X)^{-1} X^\top y. \tag{1.9}$$

La stima di massima verosimiglianza per il parametro σ^2 è invece

$$\hat{\sigma}^2 = \frac{1}{n} (y - X\hat{\beta})^\top (y - X\hat{\beta}) = \frac{1}{n} \|y - X\beta\|^2. \tag{1.10}$$

Poiché lo stimatore $\hat{\beta}$ è una trasformazione lineare di Y (normalmente distribuita), anch'esso è una variabile casuale normalmente distribuita. Pertanto

$$\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1}). \tag{1.11}$$

Si può notare che $\hat{\beta}$ è anche stima di β ai minimi quadrati in quanto minimizza rispetto a β la quantità

$$\text{argmin}\{(y - X\beta)^\top (y - X\beta)\}.$$

La differenza con la stima di massima verosimiglianza è che non viene fatta alcuna assunzione distributiva sulla risposta. Dunque la (1.9) rimane vera, però non è più possibile fare alcuna assunzione distributiva sui parametri (non è più valida la (1.11)): non esiste alcun modo per valutare l'incertezza statistica attraverso intervalli di confidenza e test statistici.
Lo stimatore di σ^2, invece, segue una distribuzione χ^2_{n-p} con $n-p$ gradi di libertà, perciò

$$n\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p}. \tag{1.12}$$

Si vuole far notare, inoltre, che lo stimatore espresso nella (1.10) non è corretto in quanto $E(n\sigma^2) = (n-2)\sigma^2 \neq n\sigma^2$.
Uno stimatore corretto sarà allora dato da

$$S^2 = \frac{n}{n-2}\hat{\sigma}^2$$

La differenza tra lo stimatore $\hat{\sigma}^2$ e S^2 è contenuta per n grande e tende ad essere infinitesima per n tendente all'infinito.

1.5.3 Residui e devianza

Si può definire il vettore dei residui come

$$e = y - X\beta = y - \hat{y}, \tag{1.13}$$

dove $\hat{y}$ è il vettore dei valori predetti dal modello. Ogni componente $e_i = y_i - \hat{y}_i$ di e può essere visto come il corrispondente campionario degli errori casuali ϵ_i. Ciò comporta che

$$\hat{\sigma}^2 = \frac{1}{n}\|y - X\beta\|^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{1.14}$$

Data la (1.13), la scomposizione ortogonale di y è pari a:

$$y^\top y = \hat{y}^\top \hat{y} + e^\top e.$$

Dunque, in un modello con intercetta la somma dei residui è pari a zero e

$$SQ_{tot} = SQ_{reg} + SQ_{res}, \tag{1.15}$$

dove:

$SQ_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$;

$SQ_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$;

$SQ_{res} = \sum_{i=1}^{n} e_i^2$.

Le componenti rappresentate nella (1.15) sono dette rispettivamente: devianza totale, devianza spiegata e devianza residua. Tale scomposizione è sfruttata dal coefficiente di determinazione R^2, utile per descrivere la bontà di adattamento del modello di regressione lineare. Tutto ciò che riguarda la bontà di adattamento dei modelli (lineari e non) sarà comunque introdotto e spiegato più approfonditamente nel Capitolo 2 *"Bontà di adattamento di modelli lineari generalizzati"*.

1.6 Modelli lineari generalizzati

1.6.1 La famiglia di dispersione esponenziale

Nel paragrafo 1.5.1 sono state introdotte le assunzioni su cui poggia un modello di regressione lineare. Molto spesso, però, è necessario trattare risposte con distribuzione diversa dalla normale e media funzione del predittore lineare (non necessariamente la funzione identità).
Si consideri, dunque, una variabile casuale Y_i, $i = 1, ..., n$, la cui funzione di probabilità dipende dai parametri (θ_i, ϕ). Si assume che la sua funzione di densità appartenga a un modello parametrico a sua volta incluso in una classe più ampia, detta famiglia di dispersione esponenziale (univariata). La distribuzione di Y_i, in particolare, appartiene a una famiglia esponenziale se può essere espressa nella forma

$$p(y_i; \theta_i, \phi) = exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{1.16}$$

con $y_i \in S \subseteq \mathbb{R}$, $\theta_i \in \Theta \subseteq \mathbb{R}$,$a_i(\phi) > 0$. Il parametro θ_i è detto parametro naturale, mentre il parametro ϕ è detto parametro di dispersione. Quando $a_i(\phi) = 1$ e $c(y_i, \phi) = c(y_i)$, allora si ha una famiglia esponenziale naturale univariata, con funzione di densità

$$p(y_i; \theta_i) = exp\left\{\theta_i y_i - b(\theta_i) + c(y_i)\right\}, \tag{1.17}$$

Se viene specificato nella (1.16) l'espressione delle funzioni $a(\cdot)$ e $b(\cdot)$ si ottiene un modello parametrico particolare. Come casi notevoli si possono trovare la distribuzione normale, gamma, binomiale, Poisson, normale multivariata e multinomiale.

1.6.2 Funzioni generatrici dei momenti e dei cumulanti, media e varianza

La funzione generatrice dei momenti di Y_i è

$$\begin{aligned} M_{Y_i}(t; \theta_i, \phi) = E(e^{tY_i}) &= \int_S e^{ty_i} p(y_i; \theta_i, \phi) dy_i \\ &= exp\{[b(\theta_i + ta_i(\phi)) - b(\theta_i)]/a_i(\phi)\}. \end{aligned} \tag{1.18}$$

La funzione generatrice dei cumulanti di Y_i risulta pertanto

$$K_{Y_i}(t; \theta_i, \phi) = [b(\theta_i + ta_i(\phi)) - b(\theta_i)]/a_i(\phi). \tag{1.19}$$

Il cumulante di ordine r, invece, è pari a

$$\kappa_r(Y_i) = \left.\frac{\partial^r K_{Y_i}(t; \theta_i, \phi)}{\partial t^r}\right|_{t=0} = a_i(\phi)^{r-1} b^{(r)}(\theta_i), \tag{1.20}$$

dove $b^{(r)}(\theta_i)$ è la derivata r-esima di $b(\theta_i)$, $r = 1, 2, ...$ Dalla (1.19) è possibile risalire a due quantità importanti quali il valore atteso e la varianza di Y_i, ottenibili, quindi, in via generale per tutti i modelli parametrici della forma (1.16). Infatti, sapendo che

$$E(Y) = \frac{d}{dt}K_Y(t)\bigg|_{t=0}, \quad Var(Y) = \frac{d^2}{dt^2}K_Y(t)\bigg|_{t=0},$$

allora

$$E(Y_i) = E_{\theta_i,\phi}(Y_i) = b'(\theta_i), \tag{1.21}$$

$$Var(Y_i) = Var_{\theta_i,\phi}(Y_i) = a_i(\phi)b''(\theta_i). \tag{1.22}$$

Si osserva che la funzione $b(\theta_i)$ determina tutti i momenti di Y_i, per tale motivo è detta generatore dei cumulanti.

1.6.3 Parametrizzazione con la media e la funzione di varianza

Posto

$$\mu_i(\theta_i) = E_{\theta_i,\phi}(Y_i), \tag{1.23}$$

dalla (1.21) si ha che

$$\mu_i(\theta_i) = b'(\theta_i). \tag{1.24}$$

Dalla (1.22) si ha invece che

$$Var_{\theta_i,\phi}(Y_i) = a_i(\phi)\mu_i'(\theta_i). \tag{1.25}$$

La (1.23) definisce una riparametrizzazione di (μ_i, ϕ). Indicato con $\theta_i(\mu_i)$ la funzione inversa che esprime θ_i in funzione di μ_i, si può esprimere la varianza attraverso

$$Var_{\mu_i,\phi}(Y_i) = a_i(\phi)b''(\theta_i)\bigg|_{\theta_i=\theta_i(\mu_i)} = a_i(\phi)v(\mu_i), \tag{1.26}$$

dove la funzione

$$v(\mu_i) = b''(\theta_i)\bigg|_{\theta_i=\theta_i(\mu_i)}, \tag{1.27}$$

è detta funzione di varianza. Tale funzione, insieme al suo dominio, caratterizza uno specifico modello della famiglia di dispersione esponenziale, pertanto, è possibile scrivere la distribuzione di Y_i nella notazione compatta

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \mu_i \in M \tag{1.28}$$

con M spazio delle medie.

1.6.4 Ipotesi e caratterizzazione di un modello lineare generalizzato

Il modello lineare classico con errori normali si basa sulle assunzioni riportate nel paragrafo 1.5.1, nel modello lineare generalizzato (*m.l.g.*), invece, tali assunzioni vengono generalizzate. In particolare:

- $Y_1, ..., Y_n$ sono variabili casuali univariate indipendenti;
- $g(E(Y_i)) = g(\mu_i) = \eta_i = \mathbf{x_i}\beta$;
- $Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i))$,

dove $g(\cdot)$ è la funzione di legame che collega μ_i a η_i, $\mathbf{x_i} = (x_{i1}, ..., x_{ip})$ costituisce il vettore delle variabili esplicative per l'i-esima unità e $\beta = (\beta_{i1}, ..., \beta_{ip})$ è il vettore dei coefficienti di regressione.
Le tre componenti che caratterizzano un modello lineare generalizzato sono quindi:

- Componente aleatoria: $Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i))$, con $Y_1, ..., Y_n$ indipendenti.
- Predittore lineare: $\eta = X\beta$
- Funzione di legame: è la funzione $g(\cdot)$ che collega μ_i al predittore lineare η_i.

Nel modello lineare normale, $g(\cdot)$ è la funzione di legame identità, perciò il predittore lineare η_i sarà uguale a μ_i.
Fra tutte le possibili funzioni di legame $g(\cdot)$ è preferita la funzione

$$g(\mu_i) = \theta_i(\mu_i),$$

poiché garantisce che il parametro naturale della famiglia esponenziale θ_i sia combinazione lineare delle covariate con coefficienti β: $\theta_i = \mathbf{x_i}\beta$, $i = 1, ..., n$. In questo caso la funzione $g(\cdot)$ è detta funzione di legame canonico.

1.6.5 Verosimiglianza nei modelli lineari generalizzati

Siano $Y_1, ..., Y_n$ variabili casuali distribuite secondo le assunzioni sopra elencate, allora $Y = (Y_1, ..., Y_n)$ ha densità congiunta data dal prodotto delle densità marginali (1.16), pertanto la funzione di log-verosimiglianza risulta pari a

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi), \tag{1.29}$$

con $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\mathbf{x_i}\beta))$.

Statistica sufficiente

Si può notare dalla (1.29) che, anche se si suppone ϕ noto, non esiste una statistica sufficiente con dimensione inferiore a n. Se tuttavia $g(\mu_i) = \theta_i(\mu_i)$, ovvero $g(\cdot)$ è la funzione di legame canonica, allora

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)} \mathbf{x_i} y_i = \left(\sum_{i=1}^{n} \frac{1}{a_i(\phi)} x_{i1} y_i, ..., \sum_{i=1}^{n} \frac{1}{a_i(\phi)} x_{ip} y_i \right) \tag{1.30}$$

è statistica sufficiente minimale per β per ogni fissato valore di ϕ, cioè la (1.30) contiene tutta l'informazione relativa ai parametri disponibile nel vettore delle osservazioni y.

Funzione di punteggio ed equazioni di verosimiglianza

Nella (1.4) è stata definita la funzione punteggio come il vettore delle derivate parziali prime della funzione di log-verosimiglianza, pertanto, in questo contesto, si può definire la funzione *score* con componenti

$$l_r = \frac{\partial l(\beta, \phi)}{\partial \beta_r} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left(y_i \frac{\partial \theta_i}{\partial \beta_r} - \frac{\partial b(\theta_i)}{\partial \beta_r} \right), \quad r = 1, ..., p \tag{1.31}$$

$$l_\phi = \frac{\partial l(\beta, \phi)}{\partial \phi} = -\sum_{i=1}^{n} \frac{a_i'(\phi)}{(a_i(\phi))^2} (y_i \theta_i - b(\theta_i)) + \sum_{i=1}^{n} c'(y_i, \phi). \tag{1.32}$$

Le equazioni di verosimiglianza per β (assumendo per il momento ϕ noto) saranno quindi pari a

$$l_r = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \beta_r} = 0, \quad r = 1, ..., p \tag{1.33}$$

Se la funzione legame è quella canonica, allora $g'(\mu_i) = 1/v(\mu_i)$, pertanto le equazioni di verosimiglianza (1.33) si semplificano nella forma

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)} y_i x_{ir} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \mu_i x_{ir}, \quad r = 1, ..., p \tag{1.34}$$

Se $a_i(\phi) = \phi$, si può notare che le equazioni si semplificano ulteriormente.
Le (1.33) possono essere scritte anche nella forma matriciale

$$(y - \mu)^\top V^{-1} D = 0, \tag{1.35}$$

con $(y - \mu)^\top = (y_1 - \mu_1, ..., y_n - \mu_n)$,

$$V = \text{diag}(Var(Y_i)), \quad i = 1, ..., n$$

e D matrice $n \times p$ con generico elemento

$$d_{ir} = \frac{\partial \mu_i}{\partial \beta_r} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{1}{g'(\mu_i)}x_{ir} \quad i = 1, ..., n, \quad r = 1, ..., p.$$

Informazione osservata e attesa

È dimostrabile che i parametri β e ϕ sono tra loro ortogonali (perciò gli stimatori di massima verosimiglianza di β e ϕ sono asintoticamente indipendenti), ne consegue che l'elemento $i_{\beta\phi}$ della matrice di informazione attesa ha elementi pari a 0. Dunque, per l'inferenza su β è sufficiente disporre del blocco d'informazione osservata o attesa relativa a β. Si ottiene

$$j_{rs} = -l_{rs} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left[\frac{\partial \mu_i}{\partial \beta_s}\frac{\partial \theta_i}{\partial \beta_r} - (y_i - \mu_i)\frac{\partial^2 \theta_i}{\partial \beta_r \partial \beta_s} \right]. \tag{1.36}$$

Se il legame è quello canonico la quantità (1.36) diventa stocastica, pertanto coinciderà con il suo valore atteso.
In generale, si ottiene informazione attesa pari a

$$i_{rs} = E[j_{rs}] = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\frac{\partial \mu_i}{\partial \beta_s}\frac{\partial \theta_i}{\partial \beta_r} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\frac{x_{ir}x_{is}}{(g'(\mu_i))^2 v(\mu_i)}, \tag{1.37}$$

che viene spesso riportata in forma matriciale

$$i_{\beta\beta} = X^\top W X, \tag{1.38}$$

dove

$$W = diag(w_i), \quad \text{con} \quad w_i = \frac{1}{(g'(\mu_i))^2 Var(Y_i)}, \quad i = 1, ..., n.$$

Definite tali quantità e richiamando il risultato generale di normalità asintotica dello SMV, ne consegue un risultato importante a livello inferenziale. Si può infatti godere dell'approssimazione

$$\hat{\beta} \dot{\sim} N(\beta, (X^\top W X)^{-1}), \tag{1.39}$$

che permetterà di valutare l'incertezza statistica tramite la costruzione di intervalli di confidenza di Wald per $\beta_r, r = 1, ..., p$ e tramite verifica d'ipotesi.

Minimi quadrati pesati iterati

Le equazioni di verosimiglianza (1.33), in genere, non ammettono soluzione esplicita. È pertanto necessario ricorrere a metodi iterativi, come il metodo di *Newton-Rapson*. Definito l_* il vettore di elementi l_r e $i = i_{\beta\beta}$ il valore atteso della matrice di informazione osservata $j_{\beta\beta}$ con elementi $-l_{rs}$, la $(m+1)$-esima iterazione fornisce l'approssimazione

$$i_{\beta\beta}(\hat{\beta}^{(m)})\hat{\beta}^{(m+1)} = i_{\beta\beta}(\hat{\beta}^{(m)})\hat{\beta}^{(m)} + l_*(\hat{\beta}^{(m)}). \tag{1.40}$$

l_* e $i_{\beta\beta}$ possono essere scritti come

$$l_* = X^\top W u, \tag{1.41}$$

$$i_{\beta\beta} = X^\top W X, \tag{1.42}$$

con $u = ((y_1-\mu_1)g'(\mu_1))^\top, ..., (y_n-\mu_n)g'(\mu_n))^\top$ e $W = diag(w_i)$, con $w_i = \frac{1}{(g'(\mu_i))^2 Var(Y_i)}$, $i = 1, ..., n$.
Pertanto, sfruttando la (1.41) e la (1.42), si può scrivere

$$X^\top W X \hat{\beta}^{(m+1)} = X^\top W z^{(m)}, \tag{1.43}$$

dove $z^{(m)} = X\hat{\beta}^{(m)} + u$.
Dopo aver inizializzato il valore di $z_i^{(0)} g(y_i)$ e di $W^{(0)}$, raggiunta la convergenza dell'algoritmo si avrà

$$\hat{\beta} = (X^\top \hat{W} X)^{-1} X^\top \hat{W} \hat{z} \tag{1.44}$$

con $\hat{z} = X\hat{\beta} + \hat{u}$

Stima del parametro di dispersione

Il problema della stima del parametro di dispersione sorge nei *glm* per risposte continue, in cui ϕ non è fissato.
Data la scarsa robustezza e l'instabilità numerica della stima di massima verosimiglianza, per la stima di ϕ si ricorre in genere al metodo dei momenti: sostituiti i valori attesi μ_i con le loro stime basate su $\hat{\beta}$, si utilizza lo stimatore con correzione

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\mu_i)}. \tag{1.45}$$

1.6.6 Devianza, bontà di adattamento e controllo del modello

Siano $Y_1, ..., Y_n$ v.c. indipendenti aventi distribuzione marginale $DE_1(\mu_i, a_i(\phi)v(\mu_i))$ e $g(\mu_i) = \mathbf{x_i}\beta$, con $a_i(\phi) = \phi/w_i$ e ϕ supposto noto per semplicità.
Si consideri, inoltre, la ripartizione di β,

$$\beta = \begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix}, \quad \text{con} \quad \beta_A = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{p_0} \end{pmatrix}, \quad \beta_B = \begin{pmatrix} \beta_{p_0+1} \\ \vdots \\ \beta_p \end{pmatrix},$$

e si supponga di voler verificare $H_0 : \beta_B = 0$ contro $H_1 : \beta_B \neq 0$.
In un modello lineare generalizzato il test rapporto di verosimiglianza suggerisce di rifiutare l'ipotesi nulla per valori elevati della statistica

$$W_P = 2(l(\hat{\beta}, \phi) - l(\hat{\beta}_0, \phi)). \tag{1.46}$$

In particolare, sotto H_0

$$W_P \sim \chi^2_{p-p_0}.$$

Si è indicata con $\hat{\beta}_0$ la stima $(\hat{\beta}_{A0}, 0)$ di β sotto H_0.
Si definisca inoltre

$$\begin{aligned} D(y; \hat{\mu}) &= 2\phi(l^M(y, \phi) - l^M(\hat{\mu}, \phi)) \\ &= 2\sum_{i=1}^{n} w_i\{y_i[\theta_i(y_i) - \theta_i(\hat{\mu}_i)] - [b(\theta_i(y_i)) - b(\theta_i(\mu_i))]\}. \end{aligned} \tag{1.47}$$

La quantità (1.47) è detta devianza (*deviance*).
La differenza $l^M(y, \phi) - l^M(\hat{\mu}, \phi)$ rappresenta una misura della diminuzione della bontà di adattamento dovuta al passaggio dal modello saturo a quello con $p < n$ variabili esplicative; si noti infatti che $l^M(y, \phi)$ rappresenta la log-verosimiglianza adattando il modello di regressione saturo con $p = n$. Vista la (1.47) si può scrivere il test del rapporto di verosimiglianza (1.46) nella forma

$$W_P = \frac{D(y; \hat{\mu}_0) - D(y; \hat{\mu})}{\phi}, \tag{1.48}$$

che segue la medesima distribuzione del test sopracitato.
Si sottolinea che, qualora ϕ sia ignoto, andrà sostituito con una sua stima consistente. La distribuzione asintotica nulla rimane $\chi^2_{p-p_0}$.

1.6.7 Criteri di informazione

Due approcci popolari alternativi al test del rapporto di verosimiglianza sono il criterio di informazione di Akaike (AIC) e il criterio di informazione Bayesiano (BIC), detto anche di Schwarz. Questi criteri, basati su penalizzazioni della log-verosimiglianza, sono infatti molto utili per confrontare due o più modelli tra loro.
Si consideri per i dati y, quindi, una successione di modelli parametrici annidati, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ... \subset \mathcal{F}_k$, con spazi parametrici corrispondenti

$$\Theta_1 \subset \Theta_2 \subset ... \subset \Theta_k \subseteq \mathbb{R}^k.$$

Si suppone quindi che il passaggio da Θ_k a Θ_{k-1} avvenga tramite l'ipotesi $\theta_k = 0$ e che le riduzioni del modello corrispondano ad analoghi annullamenti di componenti. Le

log-verosimiglianze associabili ai vari modelli sono $l(\hat{\theta}^{(1)}; y), ..., l(\hat{\theta}^{(k)}; y)$. Ovviamente la log-verosimiglianza del modello con $k-1$ parametri $l(\hat{\theta}^{(k-1)}; y)$ sarà inevitabilmente più piccola della log-verosimiglianza del modello con k parametri, per cui verrebbe selezionato sempre il modello meno parsimonioso, cioè con parametro k-dimensionale.
Senza entrare più del dovuto nell'argomento, si consideri un modello con d parametri, il valore AIC corrispondente al modello è pari a

$$AIC(\mathcal{F}_d) = 2d - 2l(\hat{\theta}^{(d)}; y), \tag{1.49}$$

pertanto la log-verosimiglianza viene penalizzata tenendo conto del numero di parametri presenti nel modello preso in considerazione.
Dato un insieme di modelli, quindi, si preferisce quello che minimizza il criterio di Akaike. Si può dimostrare, però, che questo indice seleziona tendenzialmente modelli sovra-parametrizzati, il criterio non è quindi consistente. Per risolvere tale problema si è soliti usare in alternativa il criterio BIC, basato su un approccio bayesiano. In particolare:

$$BIC(\mathcal{F}_d) = d\log n - 2l(\hat{\theta}^{(d)}; y). \tag{1.50}$$

Tale criterio è consistente. Per n elevato, tuttavia, la penalizzazione risulta eccessiva, portando alla selezione di un modello sotto-parametrizzato. Per questo motivo si è soliti usare (1.49) e (1.50) simultaneamente.
Si noti, infine, che i criteri suddetti non forniscono alcuna informazione in termini assoluti sulla qualità del modello, quindi, si possono usare esclusivamente per il confronto e la successiva selezione del modello da considerarsi migliore.

1.7 Quasi verosimiglianza

1.7.1 Modelli di quasi-verosimiglianza

Nel paragrafo 1.5 sono stati introdotti i modelli lineari in cui la distribuzione della variabile risposta Y_i è definita e pari a $Y_i \sim N(\mu_i, \sigma^2)$. Nel paragrafo 1.6, invece, sono stati introdotti i modelli lineari generalizzati, dove, come indicato nel paragrafo 1.6.4, l'assunzione di distribuzione era generalizzata ad altri casi oltre alla distribuzione normale, dunque $Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i))$.
Per aumentare ulteriormente la flessibilità e l'applicabilità della classe dei *glm* si sono introdotti i modelli di quasi-verosimiglianza (QL). Quest'ultimi sono dei modelli semi-parametrici, nei quali si specificano solamente le strutture dei primi due momenti (media e varianza) della variabile risposta Y_i e non anche una particolare forma di distribuzione. Come si può notare dalla (1.33), infatti, le equazioni di verosimiglianza per β dipendono dalla distribuzione della risposta solo attraverso $E(Y_i) = \mu_i$ e $Var(Y_i)$. Risulta quindi interessante studiare le proprietà dello stimatore $\hat{\beta}$, soluzione della (1.33), sotto le più deboli ipotesi del secondo ordine:

$$E(Y_i) = \mu_i(\boldsymbol{x}_i\beta) = g^{-1}(\boldsymbol{x}_i\beta);$$

$$Var(Y_i) = \phi v(\mu_i);$$

$$Cov(Y_i, Y_j) = 0 \quad se \quad i \neq j.$$

Il modello di quasi-verosimiglianza può essere formulato tanto per dati continui quanto per dati discreti e tiene conto di eventuale sovradispersione (o sottodispersione) dei dati, cioè si può modellare la variabile risposta Y_i anche nel caso in cui il campione $(y_1, ..., y_n)$ presenta una variabilità ampiamente superiore (o inferiore) a quella prevista dal modello adottato.

1.7.2 Inferenza basata su equazioni di stima non distorte

Con $y = (y_1, ..., y_n)$ e $\theta \in \Theta \subseteq \mathbb{R}^d$, si dice che la funzione

$$q(y;\theta) = (q_1(y;\theta), ..., q_d(y;\theta))^\top,$$

fornisce un'equazione di stima non distorta per θ se

$$E_\theta(q(Y;\theta)) = 0, \quad \forall\theta \in \Theta.$$

La non distorsione dell'equazione di verosimiglianza, alla pari della prima proprietà di Bartlett definita in (1.7), è l'assunzione principale per dimostrare la consistenza dello

stimatore di massima verosimiglianza.
Sotto le condizioni di regolarità definite nel paragrafo 1.4, è dimostrabile, applicando il teorema del limite centrale, che gli stimatori basati su $q(y;\theta)$ hanno distribuzione approssimata normale al divergere di n. Si può definire la distribuzione di $q(y;\theta)$ pari a

$$q(Y;\theta)\dot{\sim}N_d(0, J(\theta)),$$

dove $J(\theta) = E_\theta(q(Y;\theta)q(Y;\theta)^\top) = Var_\theta[q(Y;\theta)]$. Inoltre, per la legge dei grandi numeri, si può definire

$$H(\theta) = -E_\theta\left(\frac{\partial q(Y;\theta)}{\partial\theta^\top}\right) \doteq -\frac{\partial q(Y;\theta)}{\partial\theta^\top}$$

allora, indicato con $\tilde{\theta}$ la soluzione dell'equazione $q(y;\theta) = 0$:

$$\tilde{\theta} - \theta\dot{\sim}N_d(0, H(\theta)^{-1}J(\theta)H(\theta)^{-1}).$$

Riassumendo, l'assunzione parametrica $Y_i \sim DE_1(\mu_i, \phi v(\mu_i))$ può anche non essere soddisfatta, ciò che risulta essenziale è l'ipotesi sulla media e sulla varianza $v(\mu_i)$ in quanto le equazioni (1.33) per β sono equazioni di stima non distorte purché sia $E(Y_i) = \mu_i = g^{-1}(\boldsymbol{x}_i\beta)$, dunque, non è di interesse l'esplicitazione della distribuzione.

Capitolo 2

Bontà di adattamento di modelli lineari generalizzati

2.1 Introduzione

Una volta ottenuti i valori previsti $\hat{y}_i$, è d'interesse valutare quanto bene essi riescano a rappresentare i valori osservati $y_i, i = 1, ..., n$. In altre parole, risulta interessante poter valutare la bontà di adattamento del modello ai dati. A tale scopo viene utilizzato il coefficiente di determinazione R^2.

Nei modelli lineari, sfruttando la (1.15), si può definire tale coefficiente come il rapporto tra la devianza spiegata (dal modello) e la devianza totale:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = 1 - \frac{SQ_{res}}{SQ_{tot}} \tag{2.1}$$

In generale, affinché R^2 sia un buon coefficiente, devono essere rispettate le seguenti proprietà:

- $0 \leq R^2 \leq 1$;
- adimensionalità (non dipendenza dall'unità di misura delle variabili);
- indipendenza dalla numerosità n del campione;
- interpretazione facile e immediata come proporzione della devianza spiegata dal modello rispetto la devianza totale.

Si può notare che il coefficiente di determinazione in (2.1) rispetta tutte le precedenti proprietà. Quando R^2 è prossimo a 1, inoltre, il modello fornisce valori predetti prossimi a quelli osservati, pertanto l'adattamento del modello è ottimo e la distribuzione della variabile d'interesse dipende fortemente dalle variabili esplicative. Se R^2 è prossimo a 0,

invece, significa che le variabili esplicative hanno una bassa capacità di spiegare, tramite il modello, il comportamento della risposta.
Si può notare dalla (2.1) che il coefficiente di determinazione è influenzato dal numero di variabili esplicative presenti nel modello in quanto la somma dei quadrati dei residui non può aumentare e quindi R^2 non può diminuire. Si può pertanto definire una versione corretta che tenga conto del numero di parametri p presente nel modello, tale quantità è definita coefficiente di determinazione aggiustato, ed è pari a

$$R^2_{adj} = 1 - \frac{SQ_{res}/(n-p)}{SQ_{tot}/(n-1)} = R^2 - (1-R^2)\frac{p-1}{n-p} \tag{2.2}$$

Se da un lato le misure di adattamento, come il coefficiente di determinazione, sono ben stabilite per i modelli lineari, nei modelli lineari generalizzati, per cui non è semplice definire il concetto di varianza residua, non è ancora stato costruito un indice che possa spiegare la loro bontà di adattamento.
Molte sono state le proposte ma non è ancora emerso il consenso su quale sia il miglior approccio unificato in questi contesti. Scopo di tale lavoro, dunque, sarà quello di illustrare, dopo una breve panoramica di proposte precedenti, un nuovo approccio che affronta questo problema ed è universalmente applicabile, confrontandolo con le proprietà sopraelencate.

2.2 R^2 per modelli lineari generalizzati: proposte

2.2.1 R^2 basato sulla statistica rapporto di verosimiglianza: R^2_{LR} di Magee

Uno degli indici per modelli lineari generalizzati più noti è quello proposto da Magee (1990).
Si definisca innanzitutto $l(Y, \mu(\boldsymbol{X}))$ come la log-verosimiglianza del modello $E(Y|\boldsymbol{X}) = \mu(\boldsymbol{X})$ per i dati osservati Y (variabile risposta) e $\boldsymbol{X}$ (variabili esplicative). Sia invece $l(Y, \mu(\mathbf{1}_n))$ la log-verosimiglianza del modello con sola intercetta $E(Y) = \mu(\mathbf{1}_n) = \mu\mathbf{1}_n$. Tale indice fa riferimento al test rapporto di verosimiglianza, definito in (1.46), per valutare la bontà di adattamento del modello ai dati: è stato osservato, infatti, che la relazione tra il coefficiente R^2 e la verosimiglianza può essere data da

$$R^2_{LR} = 1 - exp\left[\frac{2}{n}l(Y, \hat{\mu}(\mathbf{1}_n)) - \frac{2}{n}l(Y, \hat{\mu}(\boldsymbol{X}))\right] \tag{2.3}$$

dove $\hat{\mu}(\mathbf{1}_n)$ e $\hat{\mu}(\boldsymbol{X})$ sono ottenute massimizzando le corrispondenti funzioni di verosimiglianza.
Si può notare che la quantità a esponente è negativa, al più pari a zero, pertanto $0 \leq R^2_{LR} \leq 1$. In particolare, quando $R^2 = 0$, la log-verosimiglianza spiegata dal modello con le variabili esplicative è pari a quella del modello con solo intercetta, pertanto le variabili $\boldsymbol{X}$ non danno alcuna informazione ulteriore al fenomeno studiato.

Quando $R^2 = 1$, invece, $exp\left[\frac{2}{n}l(Y,\hat{\mu}(\mathbf{1}_n)) - \frac{2}{n}l(Y,\hat{\mu}(\boldsymbol{X}))\right]$ tende a zero, cioè la log-verosimiglianza spiegata dal modello è molto maggiore rispetto a quella del modello con la sola intercetta: le variabili esplicative hanno una forte influenza sul fenomeno spiegato e il modello studiato ha un ottimo adattamento.
È facile notare che la definizione (2.3) ha le seguenti proprietà:

1. è coerente con la classica definizione di R^2 definito in (2.1) (all'aumentare della bontà di adattamento, il coefficiente di determinazione tende a 1);
2. è coerente con la definizione di massima verosimiglianza come metodo di stima, ovvero l'indice R^2 è massimizzato dalle stime di massima verosimiglianza dei parametri del modello;
3. è adimensionale, ovvero non dipende dall'unità di misura utilizzata;
4. è asintoticamente indipendente dalla dimensione n del campione;
5. può essere visto come la proporzione della variazione spiegata dal modello;
6. sia Y variabile casuale con funzione di probabilità $p(y|\beta x + \alpha)$, utilizzando lo sviluppo in serie di Taylor può essere dimostrato che, a una approssimazione del primo ordine, R^2 corrisponde al quadrato della correlazione di Pearson (1895) tra x e la funzione punteggio del modello $p(\cdot)$, cioè la derivata rispetto a β di $log[p(y|\beta x + \alpha)]$ quando $\beta = 0$;
7. è semplice dal punto di vista computazionale in quanto la massima verosimiglianza del modello è facilmente ricavabile dagli output del software R (o da qualsiasi software statistico).

L'indice precedentemente esposto può essere generalizzato a qualsiasi modello con funzione di probabilità ben definita. R^2_{LR}, in particolare, corrisponde all'indice proposto da Cox e Snell (1989) ed è una generalizzazione dell'indice per soli dati binari proposto da Maddala (1983).

2.2.2 Generalizzazione corretta di $\mathbf{R^2_{LR}}$: $\mathbf{R^2_N}$ di Nagelkerke

Si prenda a riferimento un modello di regressione logistica di cui si vuole conoscere la distribuzione della variabile risposta y. Se il modello si adatta perfettamente ai dati risulta che $l(Y,\hat{\mu}(\boldsymbol{X})) = 0$, pertanto,

$$max(R^2_{LR}) = 1 - exp\left[\frac{2}{n}l(Y,\hat{\mu}(\mathbf{1}_n))\right].$$

Quando $y = 1$ nel 50% dei casi e $y = 0$ nel restante 50% dei casi (per esempio nel caso di studi caso-controllo perfettamente bilanciati), il massimo di R^2_{LR} risulta essere pari a 0.75: $l(Y,\hat{\mu}(\mathbf{1}_n))$ è quindi un asintoto per R^2_{LR}, pertanto l'indice proposto da Magee

non risulterà mai essere pari a 1: ciò è chiaramente inaccettabile per un coefficiente di determinazione.
Per risolvere questo problema Nagelkerke (1991) ha proposto la seguente correzione,

$$R_N^2 = \left[1 - exp\left\{\frac{2}{n}l(Y, \hat{\mu}(\mathbf{1}_n)) - \frac{2}{n}l(Y, \hat{\mu}(\boldsymbol{X}))\right\}\right] \Big/ \left[1 - exp\left\{\frac{2}{n}l(Y, \hat{\mu}(\mathbf{1}_n))\right\}\right] \quad (2.4)$$

Questo indice rispetta le proprietà esposte nella sezione precedente, correggendo il difetto sopramenzionato, tuttavia è una definizione costruita *ad hoc* e sembra produrre dei risultati ingannevolmente alti, soprattutto se confrontati con l'indice R^2 ottenuto da un modello di probabilità lineare, si veda Allison (2013). Quest'ultima affermazione verrà comunque valutata adeguatamente nel capitolo successivo, dove verranno testati i coefficienti tramite simulazione.

2.2.3 R^2 basato sulla divergenza di Kullback-Leibler: R^2_{KL} di Cameron e Windmeijer

Il coefficiente di determinazione proposto da Cameron e Windmeijer (1997) si basa sulla divergenza di Kullback-Leibler (1951) per quantificare l'incertezza residua nella risposta dopo aver tenuto conto delle variabili esplicative. Per comprendere al meglio la costruzione di tale indice, dunque, risulta necessaria una breve introduzione alla divergenza di Kullback-Leibler (KL).
Vengano inizialmente considerate due densità $f(\mu_1)$ e $f(\mu_2)$, parametrizzate solo dalla media $\mu_i, i = 1, 2$. In questo caso la formula generale della divergenza di KL è

$$KL(\mu_1, \mu_2) = E_{\mu_1}\{log\left[f(y; \mu_1)/f(y; \mu_2)\right]\}. \quad (2.5)$$

Si può dunque interpretare tale indice come la misura dell'informazione persa quando $f(\mu_2)$ è usata per approssimare $f(\mu_1)$, dunque se $KL(\mu_1, \mu_2)$ è pari a zero le due quantità hanno stessa distribuzione e $f(\mu_2)$ approssima perfettamente $f(\mu_1)$.
Si deve comunque notare che la divergenza di KL non è una vera e propria distanza, infatti:

- non è rispettata la simmetria, la divergenza tra $f(\mu_1)$ e $f(\mu_2)$ è generalmente diversa da quella tra $f(\mu_2)$ e $f(\mu_1)$;
- non soddisfa la disuguaglianza triangolare.

Tuttavia è sempre positiva e pari a zero se $f(\mu_1) = f(\mu_2)$.
Se le suddette distribuzioni appartengono alla famiglia esponenziale a un parametro, cioè con $a_i(\phi) = 1$ e $c(y_i, \phi) = c(y_i)$, la divergenza di KL sarà pari a

$$KL(\mu_1, \mu_2) = 2\{(\theta_1 - \theta_2)\mu_1 - [b(\theta_1) - b(\theta_2)]\}. \quad (2.6)$$

Si consideri, successivamente, un modello di regressione. Si definisca con $l(Y, \mu(I_n))$ la log-verosimiglianza del modello di regressione saturo con $p = n$, allora è dimostrabile che

$$KL(Y, \hat{\mu}(\boldsymbol{X})) = 2l(Y, \hat{\mu}(I_n)) - 2l(Y, \hat{\mu}(\boldsymbol{X})),$$

dove $\hat{\mu}(I_n)$ indica la stima di massima verosimiglianza del corrispondente valore $\mu(I_n)$. Tale quantità può essere interpretata come la perdita di informazione che si subisce passando dal modello saturo al modello con variabili esplicative. La corrispondente quantità $KL(Y, \hat{\mu}(\mathbf{1}_n))$ sarà quindi interpretabile come la differenza tra l'informazione fornita dal modello saturo rispetto a quella data dal modello con sola intercetta.
Fornite le suddette quantità, l'indice di Cameron-Windmeijer è definito come la riduzione dell'informazione ottenuta adattando il modello di regressione:

$$R^2_{KL} = 1 - \hat{KL}(Y, \hat{\mu}(\boldsymbol{X}))/\hat{KL}(Y, \hat{\mu}(\mathbf{1}_n)), \tag{2.7}$$

dove con $\hat{KL}(\cdot)$ si indica la stima della corrispondente quantità $KL(\cdot)$. Poiché entrambe le quantità in (2.7) sono deviazioni, R^2_{KL} può essere interpretato come il rapporto di riduzione della devianza dovuto alle variabili esplicative $\boldsymbol{X}$. In altre parole, se tale quantità risulta pari a 0, significa che $\hat{KL}(Y, \hat{\mu}(\boldsymbol{X}))$ è uguale a $\hat{KL}(Y, \hat{\mu}(\mathbf{1}_n))$, pertanto la differenza di informazione tra modello saturo e modello con variabili esplicative risulta pari alla differenza di informazione tra modello saturo e modello con sola intercetta, dunque si può affermare che $\boldsymbol{X}$ non da alcuna informazione ulteriore e il modello di regressione preso a riferimento non ha un buon adattamento.
Per modelli di regressione basati sulla densità (1.16), il coefficiente di regressione definito in (2.7) gode delle seguenti proprietà:

- non diminuisce con l'aggiunta di variabili esplicative;
- $0 \leq R^2_{KL} \leq 1$;
- se si adotta la funzione di legame canonico nel modello, R^2_{KL} può essere interpretato come la frazione di incertezza spiegata dal modello utilizzato.
- è uguale all'indice rapporto di verosimiglianza $1 - l(Y, \hat{\mu}(\boldsymbol{X}))/l(Y, \hat{\mu}(\mathbf{1}_n))$ se e solo se $l(Y, \hat{\mu}(I_n)) = 0$;

Si sottolinea che l'ultima proprietà è di particolare interesse in quanto l'indice rapporto di verosimiglianza, che misura la riduzione in termini di log-verosimiglianza dovuta all'inclusione delle variabili esplicative nel modello, è spesso usata come un pseudo coefficiente di determinazione R^2.
Come nel caso lineare, R^2_{KL} è influenzato anche dal numero di parametri p presenti nel modello, si potrebbe quindi osservare un aumento della bontà di adattamento anche

dopo aver inserito delle variabili esplicative irrilevanti per il fenomeno studiato. È quindi necessario tenerne conto: essendo la (2.7) basata su delle devianze, è possibile risolvere il problema dividendo le devianze stesse per i corrispondenti gradi di libertà. La versione aggiustata sarà pertanto pari a

$$R^2_{KL,adj} = 1 - \frac{\hat{KL}(Y, \hat{\mu}(\boldsymbol{X}))/(n-p)}{\hat{KL}(Y, \hat{\mu}(\mathbf{1}_n))/(n-1)}. \tag{2.8}$$

2.2.4 R^2 basato sulla funzione di varianza: R^2_V di Dabao Zhang

Tutti i coefficienti di determinazione sopramenzionati forniscono una buona misura della bontà di adattamento ai dati, tuttavia possiedono un grosso difetto: per determinarli la funzione di probabilità deve essere completamente specificata.
Nei casi in cui, per esempio, è specificata solo la funzione media e varianza (per esempio nei modelli di quasi-verosimiglianza, si veda paragrafo 1.7), la funzione di log-verosimiglianza risulta ignota, pertanto non è più possibile ottenere una stima di tali indici.
Il coefficiente R^2 proposto da Zhang (2017) si basa, invece, su una misura più semplice di incertezza: la funzione di varianza. Rifacendosi alla (1.26), si ponga $a_i(\phi) = \phi$ e si supponga nota la funzione di varianza $v(\cdot)$. Quindi, posto $E(y_i|X_i) = \mu(X_i), i = 1, ..., n$,

$$var(y_i|X_i) = \phi v(\mu(X_i)). \tag{2.9}$$

Sulla base della (2.9), quindi, è possibile affermare che, fintantoché la media $\mu(X_i)$ può essere modellata e collegata adeguatamente a un set di variabili esplicative, un modello lineare generalizzato, con funzione di varianza $v(\cdot)$ nota, può essere studiato per valutare l'utilità delle variabili esplicative.
Se da un lato la funzione di varianza descrive l'effetto della media sulla variazione della variabile risposta, dall'altro è dimostrabile che, insieme al suo dominio, può caratterizzare uno specifico modello della famiglia esponenziale, cioè si può specificare la distribuzione di Y_i nella forma (1.28).
Per una variabile risposta con media che varia da a a b, la sua variazione si muove conseguentemente lungo la funzione di varianza da $\phi v(a)$ a $\phi v(b)$. Invece di usare la classica formula della distanza euclidea $(a-b)^2$, quindi, per misurare la variazione della variabile risposta si può usare

$$d_V(a,b) = \left\{ \int_a^b \sqrt{1 + [v'(t)]^2} dt \right\}^2 \tag{2.10}$$

Come ci si può aspettare, nel caso di funzioni di varianza non lineari (si veda, per esempio, la distribuzione binomiale), la distanza euclidea può differire notevolmente dalla distanza (2.10).

Data la specificazione della varianza in (1.27), come dimostrato da Morris (1982), si può definire, per le famiglie esponenziali più popolari, una formula generale della varianza che, assunto $\delta_2 \neq 0$, è pari a:

$$v(\mu) = \delta_2\mu^2 + \delta_1\mu + \delta_0. \tag{2.11}$$

Allora è possibile determinare $d_V(a, b)$ tramite

$$d_V(a,b) = \frac{1}{16\delta_2^2}\left\{log\frac{v'(b) + \sqrt{1 + [v'(b)]^2}}{v'(a) + \sqrt{1 + [v'(a)]^2}} + v'(b)\sqrt{1 + [v'(b)]^2} - v'(a)\sqrt{1 + [v'(a)]^2}\right\}^2. \tag{2.12}$$

Quando $\delta_2 = 0$, invece, la funzione di varianza risulta lineare o costante (come nel caso della distribuzione di Poisson) e pertanto

$$d_V(a,b) = (1 + \delta_1^2)(b - a)^2.$$

Per specificare un coefficiente che determini la bontà di adattamento del modello ai dati, quindi, è possibile usare la distanza (2.12). In particolare, la variazione totale della variabile risposta Y è pari a $\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{1}_n))$ mentre il modello con variabili esplicative $\boldsymbol{X}$ riduce la varianza non spiegata in Y a $\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\boldsymbol{X}))$. Si può pertanto definire il coefficiente di determinazione tramite

$$R_V^2 = 1 - \frac{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\boldsymbol{X}))}{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{1}_n))} \tag{2.13}$$

Si può notare, come precedentemente introdotto, che tale coefficiente è specificato utilizzando solo la funzione di media e varianza, pertanto, $\mu_i(\boldsymbol{X})$ e $\mu_i(\mathbf{1}_n)$ possono essere derivate tramite stimatori di quasi-verosimiglianza, diversi quindi dai classici stimatori di massima verosimiglianza.
Siccome $v'(\cdot)$ è costante per la funzione di distribuzione normale e Poisson, R_V^2 risulta coerente con la classica definizione (2.1), quindi, nel caso di modelli di regressione lineari che seguono la distribuzione normale e per modelli log-lineari che seguono la distribuzione di Poisson, risulta $R_V^2 = R^2$.
Come per il coefficiente R_{KL}^2, anche R_V^2 è influenzato dal numero di parametri p presenti nel modello. È quindi necessario tenere conto delle variabili esplicative inserite nel modello, perciò, analogamente alla (2.2) e alla (2.8), si definisce il coefficiente di determinazione R_V^2 aggiustato come

$$R_{V,adj}^2 = 1 - \frac{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\boldsymbol{X}))/(n-p)}{\sum_{i=1}^n d_V(y_i, \hat{\mu}_i(\mathbf{1}_n))/(n-1)}. \tag{2.14}$$

Dalla costruzione della (2.14) si può notare che, anche in questo caso, come per il coefficiente non aggiustato, tale indice è ben definito specificando solo la funzione di

media e varianza.
Eventuali pregi e difetti di questo coefficiente di determinazione saranno valutati nel capitolo successivo, dove, tramite simulazione, verrà testato "sul campo", applicandolo a diverse distribuzioni fissate.

Capitolo 3

Studio empirico

3.1 Software e simulazione

Al fine di valutare le performance degli indici R^2, sono state condotte due diverse simulazioni, la prima per confrontare tra loro gli indici non aggiustati e la seconda per valutare gli indici aggiustati secondo i gradi di libertà, al variare del numero p di parametri.
Nel primo caso è stato generato un totale di 100 *data-set* secondo un parametro β fissato, che verrà fatto variare da 0 a 6 con *steps* di 0.1. Ogni *data-set* è composto da 50 osservazioni provenienti da distribuzioni della stessa famiglia esponenziale ma con medie μ_1 e μ_2 differenti: 25 osservazioni, quindi, provengono dalla prima popolazione di media μ_1, le restanti 25 provengono dalla seconda popolazione di media μ_2. In seguito, verrà usata la variabile X_1 per indicare le due differenti popolazioni. Al termine della simulazione, i coefficienti di determinazione verranno calcolati per ogni parametro fissato β come media degli indici calcolati in ogni *data-set* dei 100 simulati, pertanto il j-esimo indice, corrispondente al j-esimo valore di β, sarà pari a:

$$R^2_{ij} = \frac{1}{100} \sum_{i=1}^{100} Rij$$

con $i = 1, ..., 100$ e j fissato variante secondo β.
Si indichi con η il prodotto tra le variabili esplicative e il parametro β fissato, cioè $\eta = X\beta$, le variabili risposta Y sono state simulate, quindi, da tre diverse famiglie esponenziali:

- *Distribuzione binomiale:* per mezzo del parametro β fissato, sono state trovate le funzioni media μ_1 e μ_2 tramite l'inversa della funzione *logit* (funzione di legame canonica nel modello binomiale), pertanto: $\mu_1 = \frac{e^{-\beta}}{1+e^{-\beta}}$ e $\mu_2 = \frac{e^{\beta}}{1+e^{\beta}}$. In forma generale, dunque, risulterà $\mu(X_1) = \frac{e^{X_1\beta}}{1+e^{X_1\beta}} = \frac{e^{\eta}}{1+e^{\eta}}$, dove con X_1 pari a 1 o a -1 si è indicata la variabile indicatrice delle due differenti popolazioni;

- *Distribuzione Poisson:* per la distribuzione Poisson si è presa come riferimento la funzione di legame canonico $log(\cdot)$, pertanto la media, in generale, sarà pari a $\mu(X_1) = e^{\eta}$;
- *Distribuzione Gamma:* nel caso della distribuzione Gamma, costituita da due parametri, si è fissato il parametro di forma $\nu = 100$ e generato il parametro di scala λ_i in funzione di X_1 tramite $\lambda_i(X_1) = \frac{0.01}{\beta_i + 1 + \eta}$

Per la valutazione dei coefficienti di determinazione aggiustati secondo i gradi di libertà definiti nella (2.8) e nella (2.14), si è ripetuto un procedimento del tutto simile, ciò che variava, in questo caso, era il numero di variabili esplicative prese a riferimento. Nel primo caso, quindi, si sono calcolate le medie delle due popolazioni tramite $\mu(X_1; \beta)$, cioè $\eta = X_1\beta$. Nel secondo caso $\eta = X_{12}\beta$, dove con X_{12} si è indicata la matrice contente nella prima colonna la variabile X_1 e nella seconda la variabile X_2 generata da una distribuzione normale. Nel terzo caso, infine, $\eta = X_{123}\beta$, dove con X_{123} si indica la matrice contenente nella prima colonna X_1, nella seconda X_2 e nella terza X_3, variabile esplicativa generata da una distribuzione normale, indipendente da X_2. Si sottolinea che nel modello influisce sulla variabile risposta solo la prima covariata X_1 tramite il coefficiente β. Per attuare tale procedura, in particolare, si è utilizzata la libreria `rsq`. Scopo di questa simulazione sarà quello di confrontare i diversi indici al variare del numero p di parametri presenti nel modello, per valutare se effettivamente colgono il decremento dei gradi di libertà nel modello.

3.2 Risultati

Coefficienti di determinazione non aggiustati

Le prime osservazioni sugli indici simulati sono state fatte agli estremi, cioè, in corrispondenza del minimo e del massimo valore del parametro β. Quando $\beta = 0$, in particolare, ci si aspetta che $\mu_1 = \mu_2$, perciò i corrispondenti coefficienti di determinazione dovrebbero essere tutti pari a zero. Per β tendente all'infinito, invece, ci si aspetta che $|\mu_2 - \mu_1|$ raggiunga il suo massimo. Ciò significa che le due popolazioni sono ben distinte e che quindi la variabile X_1 discrimina bene le due popolazioni, ci si aspetta pertanto che il coefficiente di determinazione tenda al valore unitario.
Nella *Figura 3.1* si possono osservare i coefficienti di determinazione calcolati per il modello binomiale. Come già dimostrato da Nagelkerke (1991), si nota che R^2_{LR}, a differenza degli altri indici, al crescere di β tende a un valore diverso da 1, pari a 0.75. Gli altri indici studiati tendono tutti al valore unitario, come atteso. Complessivamente, l'indice R^2_N, per ogni valore di β, fornisce risultati maggiori rispetto agli altri, mentre l'indice d'interesse R^2_V si colloca nella zona intermedia fra R^2_N e R^2_{KL} che fornisce, in generale, i valori più bassi. Quando $\beta = 0$, invece, tutti i coefficienti tendono ad oscillare intorno al valore zero, come atteso.

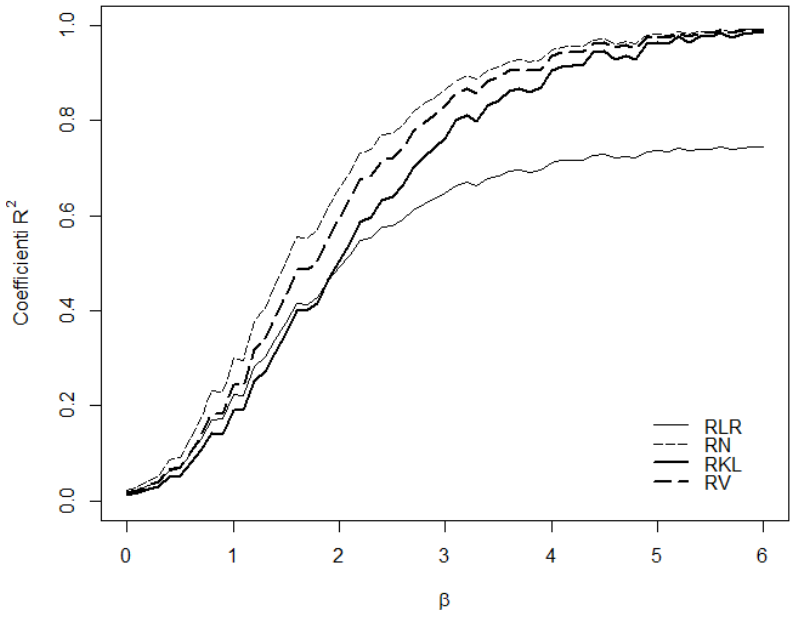

FIGURA 3.1: Coefficienti di determinazione nei modelli binomiali.

Nella *Figura 3.2* e nella *Figura 3.3* si possono osservare, invece, i coefficienti di determinazione calcolati nei modelli Poisson e Gamma rispettivamente.

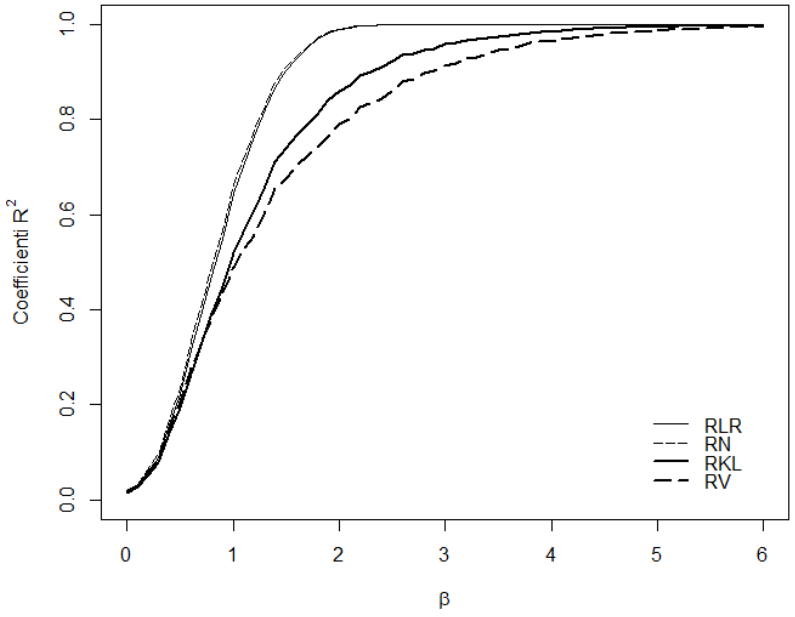

FIGURA 3.2: Coefficienti di determinazione nei modelli Poisson.

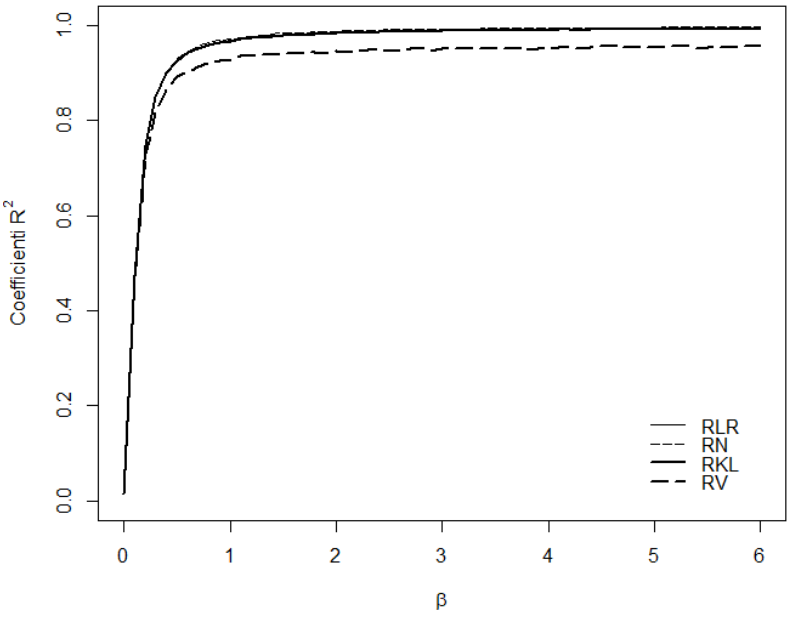

FIGURA 3.3: Coefficienti di determinazione nei modelli Gamma.

A differenza del modello binomiale, nel modello Poisson l'indice R^2_V produce i valori più bassi, messo a confronto con gli altri coefficienti di determinazione calcolati. Si osserva, inoltre, che sia R^2_{LR} che R^2_N approcciano a 1 molto velocemente (la bontà di adattamento del modello fornita potrebbe risultare quindi molto più grande rispetto la realtà), R^2_{KL} tende ad avere andamento e valori simili all'indice R^2_V. Tutti i coefficienti di determinazione, comunque, sono vicini allo zero quando $\beta = 0$, mentre tendono al valore unitario al crescere di $|\mu_2 - \mu_1|$.

Per il modello Gamma si producono dei coefficienti di determinazione simili tra loro che tendono molto velocemente al valore unitario. Vista la velocità con cui i coefficienti tendono a 1, è lecito supporre che siano inclini a sovrastimare la bontà di adattamento del modello.

In generale si può notare un comportamento simile da parte dei coefficienti di determinazione: tutti, più o meno velocemente, crescono con il crescere di β e tendono tutti al valore unitario. R^2_N, inoltre, suggerisce in tutti e tre i casi un'interpretazione "ottimistica" della proporzione di variazione espressa dal modello, esagerando frequentemente i valori del coefficiente.

Spostando l'interesse esclusivamente all'indice R^2_V, si può osservare che, nel caso di un modello binomiale, sembra essere un giusto compromesso fra i coefficienti di determinazione precedenti, in quanto, come già notato, assume valori intermedi rispetto agli altri indici. Nel modello Poisson, invece, vengono assunti i valori più bassi, confrontati con gli altri coefficienti, ciò può portare a pensare che, in questo caso, il coefficiente R^2_V

possa sottostimare la reale bontà di adattamento del modello, lo stesso avviene per il modello Gamma.
Visti i risultati ottenuti, dunque, si può affermare che i pregi del coefficiente di interesse R_V^2 sono:

- $0 \leq R_V^2 \leq 1$;
- non diminuisce con l'aggiunta di variabili esplicative nel modello;
- non tende a sovrastimare la proporzione di variazione spiegata dal modello;
- può essere adattato anche per modelli la cui distribuzione non è specificata.

Per valutare al meglio l'ultima qualità individuata, si adatterà nel capitolo successivo un modello di quasi verosimiglianza la cui distribuzione non è specificata: saranno specificate solamente la funzione di media e varianza, i soli dati necessari per il calcolo dell'indice d'interesse in questo lavoro.

Coefficienti di determinazione aggiustati

Si valuteranno ora i due indici aggiustati definiti nel capitolo precedente.
Dalla *Figura 3.4* si possono osservare le differenze dei coefficienti, sia non aggiustati che aggiustati, calcolate regredendo Y vs (X_1, X_2, X_3) e Y vs X_1, dove X_2 e X_3 sono delle variabili casuali con distribuzione normale standard, indipendenti tra loro. Si nota che sia per R_{KL}^2 che per R_V^2 i coefficienti non aggiustati risultano essere sempre maggiori del rispettivo coefficiente aggiustato all'aumentare di β, a conferma dell'attendibilità della correzione usata. Si nota inoltre, come ci si può aspettare, che le differenze tendono a zero per β tendente all'infinito: più un coefficiente del modello è significativo più il predittore corrispondente ha influenza sul fenomeno studiato, pertanto X_1, in questo caso, avrà sempre più importanza per il fenomeno al crescere di β e la correzione attuata avrà sempre meno rilevanza sul coefficiente (si ricordi che nel modello influisce sulla variabile risposta solo la prima covariata X_1 tramite il coefficiente β).

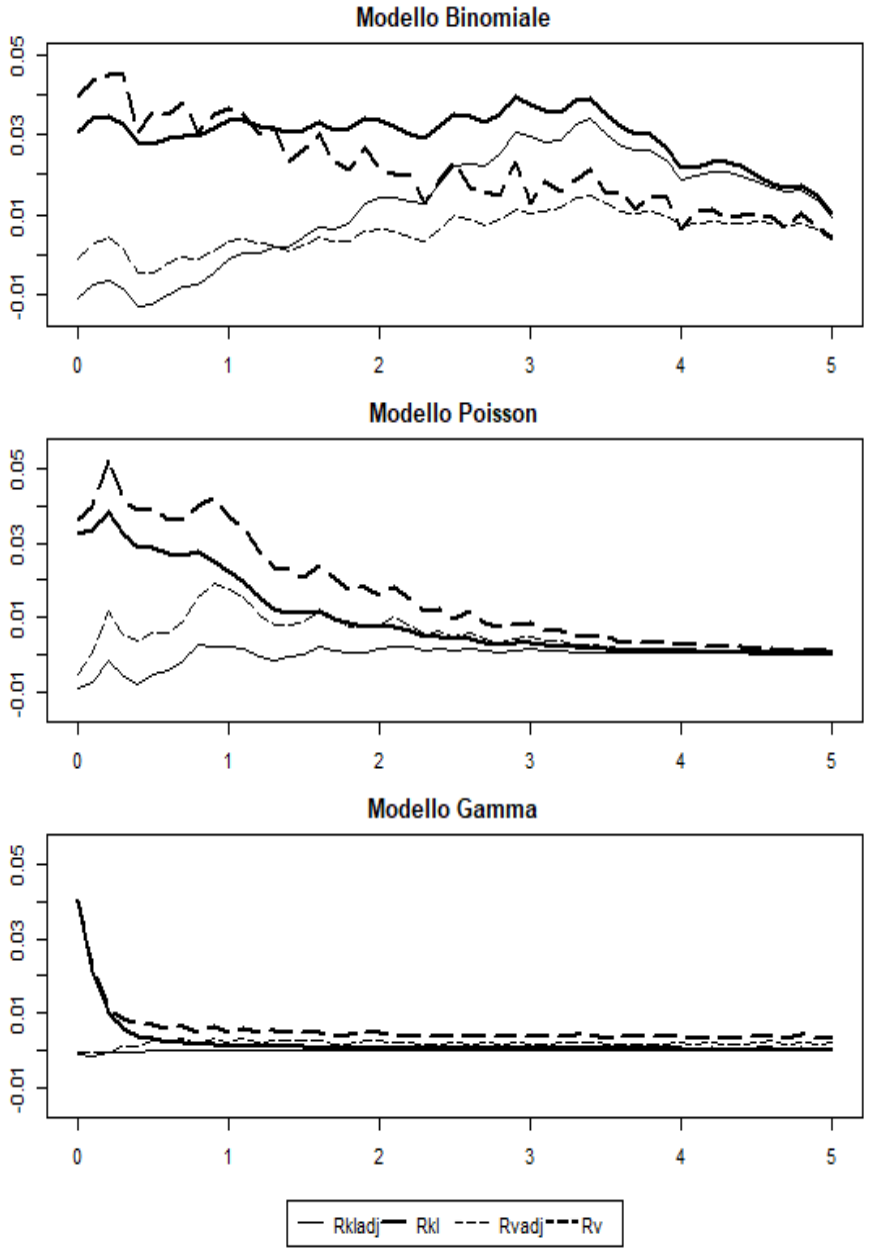

FIGURA 3.4: Differenze dei coefficienti di determinazione calcolate regredendo Y vs (X_1, X_2, X_3) e Y vs X_1 (nell'asse delle ascisse β, nell'asse delle ordinate $R^2(X_1, X_2, X_3) - R^2(X_1)$).

Capitolo 4

Applicazione su dati reali

Obiettivo

In questo capitolo si illustreranno le diverse definizioni di R^2 descritte nei capitoli precedenti, applicandoli a dei *data-set* reali. In particolare, verranno testati rispettivamente su un modello binomiale, su un modello per dati di conteggio (Poisson) e su un modello di quasi verosimiglianza, selezionato appositamente per testare il solo coefficiente R_V^2 con lo scopo di verificare la sua importante qualità: dare una misura della bontà di adattamento del modello senza specificazione di alcuna distribuzione.

4.1 Analisi sull'insolvenza della clientela

I dati contenuti nel *data-set* messo a disposizione da Fahrmeir e Tutz (2001), sono relativi a un'indagine effettuata da una banca tedesca su 1000 clienti. La variabile risposta Y_i di interesse è una variabile dicotomica associata dalla banca al cliente che ne descrive la sua situazione: `mal` se il cliente è insolvente e `buen` altrimenti. Le variabili esplicative raccolte sono:

- stato del conto corrente (non ha un conto corrente, cattivo conto corrente, buon conto corrente)*(C)*;
- durata del credito in mesi *(M)*;
- pagamento di crediti precedenti (è stato un cattivo creditore oppure è stato un buon creditore) *(P)*;
- uso previsto del credito (privato o professionale) *(U)*;
- ammontare del credito richiesto *(D)*;
- genere del richiedente *(S)*;
- stato civile (vive solo oppure non vive solo) *(E)*.

L'interesse di questa analisi era di valutare se la probabilità di insolvenza è influenzata dalle variabili esplicative indicate in precedenza (dette fattori di rischio).
Si vuole far notare che lo studio condotto è uno studio caso-controllo, dove le variabili risposta sono fissate e le esplicative vengono raccolte retrospettivamente. L'unico modello che permette di modellare la variabile risposta in funzione delle esplicative in questa tipologia di studio è il modello di regressione logistica, che verrà pertanto selezionato.
Ogni variabile qualitativa di k modalità è stata codificata con $k-1$ variabili *dummy*.
Vengono considerati diversi modelli possibili (senza considerare eventuali interazioni). Per ogni modello valutato si riportano in *Tabella 4.1* i valori dei coefficienti di determinazione per modelli lineari generalizzati, *AIC* (criterio di informazione di Akaike) e *BIC* (criterio di informazione di Schwarz). Scopo di questo procedimento è capire se effettivamente le informazioni date dagli indici rappresentati sono in concordanza con la selezione *backward* effettuata in parallelo tramite test log-rapporto di verosimiglianza (coefficienti di determinazione più alti implicano un modello migliore).

	R^2_{LR}	R^2_N	R^2_{KL}	R^2_{KLadj}	R^2_V	R^2_{Vadj}	AIC	BIC
(C,M,P,U,D,S,E)	0.1849	0.2621	0.1673	0.1606	0.1980	0.1915	1035.3	1079.5
(C,M,P,U,S,E)	0.1841	0.2610	0.1665	0.1606	0.1971	0.1915	1034.3	1073.6
(C,M,P,U,E)	0.1834	0.2600	0.1659	0.1608	0.1965	0.1916	1033.1	1067.5
(C,M,P,U,S)	0.1814	0.2572	0.1638	0.1588	0.1936	0.1888	1035.6	1070.0
(C,M,P,S,E)	0.1768	0.2506	0.1592	0.1541	0.1892	0.1843	1041.2	1075.6
(C,M,P,E)	0.1764	0.2501	0.1588	0.1546	0.1888	0.1847	1039.7	1069.2
(C,M,P,U)	0.1742	0.2470	0.1567	0.1524	0.1854	0.1813	1042.3	1071.8
(C,M,P)	0.1680	0.2383	0.1506	0.1472	0.1787	0.1754	1047.8	1072.3

Tabella 4.1: Analisi degli indici relativi al data-set sull'insolvenza della clientela.

Dalla *Tabella 4.1* si può osservare quanto notato nel capitolo precedente attraverso la simulazione: in genere il coefficiente R^2_N (di Nagelkerke) assume i valori più alti rispetto agli altri, R^2_V valori intermedi tra R^2_N e R^2_{LR} e R^2_{KL} i valori più bassi, l'ordine è quindi rispettato.
In grassetto è stato evidenziato il modello che la selezione *backward* ha selezionato. Si può notare che i coefficienti di determinazione non aggiustati, come è doveroso aspettarsi, propendono per il modello con il maggior numero di variabili, cioè con $p = 7$. Ciò che invece risulta più interessante notare è la concordanza degli R^2 aggiustati con i criteri *AIC*, *BIC* e, soprattutto, con il test log-rapporto di verosimiglianza usato per la selezione del modello. Tutte queste caratteristiche, usuali per la definizione di coefficiente di determinazione, sono rispettate anche dall'indice R^2_V e R^2_{Vadj}.
Sembra quindi lecito affermare che, per il modello binomiale, quanto osservato tramite simulazione è rispettato anche in *data-set* reali. I coefficienti aggiustati, inoltre, danno

gli esiti aspettati, cioè colgono adeguatamente il numero p di parametri presenti nel modello.

4.2 Analisi della riproduzione dei limuli

In questo capitolo viene analizzato il *data-set* messo a disposizione da Agresti (1996) riguardante uno studio sul comportamento riproduttivo del limulo (un artropode marino diffuso lungo la costa Atlantica).
La variabile risposta Y_i è il numero di satelliti (numero di maschi che contribuiscono alla fecondazione delle uova), di interesse sarà valutare se alcune caratteristiche della femmina influenzano il numero di satelliti che essa riceve. In particolare, le variabili concomitanti raccolte sono:

- peso della femmina in *kg (W)*;
- larghezza del carapace in *cm (CW)*;
- colore della femmina (chiaro, medio, abbastanza scuro, scuro) *(C)*;
- condizione delle spine (entrambe in buono stato, una solo danneggiata, entrambe danneggiate) *(SC)*.

Come nel paragrafo precedente sono stati considerati diversi modelli al fine di valutare gli indici presentati. In questo studio sono stati considerati, inoltre, anche i termini quadratici per il peso e la larghezza del carapace. In *Tabella 4.2* sono pertanto riportati i valori di interesse (in grassetto il modello selezionato dalla procedura *backward*: le variabili selezionate sono il peso, la larghezza e il termine quadratico della larghezza). Si sottolinea che, dove non indicato esplicitamente, se le variabili W e CW sono selezionate nel modello allora si considera anche il loro termine quadratico, altrimenti non si considera nessuno dei due termini.

	R^2_{LR}	R^2_N	R^2_{KL}	R^2_{KLadj}	R^2_V	R^2_{Vadj}	AIC	BIC
(W,CW,C,SC)	0.4309	0.4324	0.1541	0.1074	0.1515	0.1046	910.6	942.1
(W,CW,C)	0.4298	0.4313	0.1536	0.1177	0.1518	0.1158	906.9	932.1
(W,CW)	0.4057	0.4070	0.1423	0.1218	0.1422	0.1218	908.1	923.8
(CW,C)	0.3786	0.3798	0.1301	0.1040	0.1264	0.1002	917.8	936.7
(W,C)	0.4153	0.4167	0.1467	0.1212	0.1417	0.1160	907.2	926.2
(W,CW,−W²)	0.4047	0.4060	0.1418	0.1266	0.1422	0.1270	906.4	919.0

Tabella 4.2: Analisi degli indici relativi al data-set sulla riproduzione dei limuli.

Anche in questo caso i coefficienti aggiustati colgono il modello selezionato fornendo il valore più elevato fra quelli calcolati nei diversi modelli rappresentati. I coefficienti,

inoltre, sono coerenti con la classica definizione di R^2, si può infatti notare che al crescere del numero di parametri nel modello i coefficienti di determinazione tendono a crescere, così non è per quelli aggiustati, che riescono a cogliere correttamente il numero p di parametri nel modello.
Una differenza notevole avviene fra gli indici R^2_{LR} e R^2_N contro R^2_{KL} e R^2_V: per il modello selezionato, per esempio, si passa da un coefficiente di circa 0.4 per i primi due a un coefficiente di circa 0.14 per i secondi due. Si è portati pertanto ad affermare che i coefficienti proposti da Magee e da Nagelkerke tendono a sovrastimare gravemente la proporzione di variazione spiegata dai modelli di Poisson (per dati di conteggio).

4.3 Analisi di uno studio di teratologia sui ratti

I dati contenuti in questo *data-set* fanno riferimento ad un esperimento di laboratorio volto a studiare l'effetto del regime alimentare sullo sviluppo fetale dei ratti. L'esperimento consisteva nell'assegnare casualmente le 58 femmine di ratto, rese gravide, a uno di quattro gruppi (al primo gruppo veniva iniettato placebo, agli altri tre supplementi diversi di ferro). Alla fine dell'esperimento i ratti venivano sacrificati e si misurava:

- il numero di feti morti nella nidiata;
- numero totale di feti nella nidiata;
- livello emoglobina della madre *(h)*;
- gruppo di appartenenza della madre *(g)*.

Ciò che si vuole studiare, dunque, è il rapporto tra il numero di feti morti sul totale, perciò la probabilità di morte date le variabili esplicative misurate. Si evidenzia che, a causa di covariate non rilevate e della variabilità genetica, la probabilità di morte dei feti può variare, sia all'interno della stessa nidiata, o all'interno dello stesso gruppo o, infine, tra nidiate con stesso livello di emoglobina della madre.
Inizialmente viene adattato un modello binomiale (l'interesse è infatti una probabilità). Sfruttando la (1.45) si può trovare una stima del parametro di dispersione che, secondo un modello binomiale, dovrebbe essere pari a 1. La stima che ne risulta è $\tilde{\phi} = 2.906$ a confermare una chiara sovradispersione dei dati analizzati. Si decide quindi di adattare un modello di quasi-verosimiglianza: si potrà così mettere alla prova il coefficiente di determinazione R^2_V in quanto non viene definita alcuna distribuzione ma solo la funzione media e varianza.
Si considera quindi inizialmente il modello binomiale (nel quale si è considerato anche il termine di interazione) e successivamente il modello di quasi- verosimiglianza riportando in *Tabella 4.3* i valori dei coefficienti di determinazione R^2_V e R^2_{Vadj} (da notare che nemmeno i criteri AIC e BIC sono calcolabili per il modello di quasi-verosimiglianza in quanto si basano sulla funzione di log-verosimiglianza).

	R^2_{LR}	R^2_N	R^2_{KL}	R^2_{KLadj}	R^2_V	R^2_{Vadj}	AIC	BIC
Modello binomiale								
*(g,h,g*h)*	0.9975	0.9975	0.6809	0.6632	0.7666	0.7537	242.0	250.3
Modello di quasi-verosimiglianza								
*(g,h,g*h)*					0.7666	0.7537		
(g,h)					0.7365	0.7270		
(g)					0.7187	0.7138		
(h)					0.7056	0.7004		

TABELLA 4.3: Analisi degli indici relativi al data-set sullo studio di teratologia sui ratti.

A dimostrazione di quanto affermato precedentemente, si possono osservare, per esempio, i coefficienti R^2_{LR} e R^2_N calcolati nel modello con distribuzione binomiale: viene affermato che quasi il 100% dei dati sono spiegati dal modello adottato, il che è evidentemente impossibile. Come già notato, quindi, si può affermare che, a causa di sovradispersione, i risultati sono inevitabilmente ingannevoli.

Adattato un modello di quasi-verosimiglianza si osservano i coefficienti di determinazione R^2_V: suggeriscono di assumere il modello più complesso tra i quattro, cioè quello con termine di interazione: oltre all'effetto delle singole variabili, esiste un effetto del gruppo sul livello di emoglobina della madre, cioè, ratti appartenenti a gruppi diversi tendono ad avere livelli di emoglobina diversi.

Capitolo 5

Conclusioni

Lo scopo di questo lavoro era quello di descrivere e verificare il lavoro condotto da Zhang (2017), riguardante una nuova definizione di R^2 per modelli lineari generalizzati, indicato con R^2_V per la sua forte relazione con la funzione di varianza definibile sia per modelli con distribuzione nota (binomiale, Poisson, Gamma,...) sia per modelli la cui distribuzione non è specificata (modelli di quasi-verosimiglianza).
Inizialmente è stata esposta la teoria che sta alla base sia del coefficiente di interesse R^2_V sia dei coefficienti esposti precedentemente da altri autori e basati su termini diversi della distribuzione. Nei due capitoli successivi si sono messe alla prova tali definizioni: R^2_{LR}, oltre a sovrastimare la variazione spiegata dal modello, ha il grosso problema di tendere a un valore diverso da 1 per β tendente a infinito quando la distribuzione è quella binomiale. L'indice R^2_N corregge adeguatamente questo problema ma anch'esso sembra sovrastimare notevolmente la bontà di adattamento del modello ai dati. L'indice R^2_{KL} sembra essere un buon coefficiente in quanto si basa sulla devianza e fornisce risultati coerenti se confrontato con gli altri indici.
Tutti e tre i coefficienti, però, si basano sulle ipotesi sottostanti la distribuzione adottata: frequentemente le imposizioni imposte dal modello non sono rispettate nella pratica, si veda, per esempio, il caso di sovradispersione (o sottodispersione), dunque i coefficienti risultano inevitabilmente distorti e forniscono un risultato diverso dalla realtà. Il punto di forza di R^2_V è appunto quello di basarsi sulla funzione di varianza che viene colta perfettamente, in quanto non viene fatta alcuna ipotesi distributiva alla base di questo coefficiente. Dunque, non solo R^2_V può essere adattato per modelli la cui distribuzione non è specificata, ma coglie eventuali violazioni delle ipotesi distributive.
Vista, inoltre, l'applicazione a *data-set* reali è lecito affermare che, oltre a dare una corretta idea della bontà di adattamento, può essere anche usato come valido indice per effettuare la selezione del modello, in quanto i risultati sono coerenti con la selezione *backward*.
In conclusione, si può affermare che le qualità osservate durante il corso del lavoro fanno di R^2_V una valida alternativa del corrispondente R^2 per modelli lineari.

Appendice A

Procedura di simulazione

Listing A.1: Implementazione dei coefficienti di determinazione

```
#CREAZIONE R QUADRO DI MAGEE: Rlr
Rlr = function(fit,distr){
  fam = distr
  dati = fit$model
  n = fit$df.null+1
  fit0 = update(fit,.~1,family=fam,data=dati)
  value = 1 - exp((2/n)*(as.numeric(logLik(fit0))
  -as.numeric(logLik(fit))))
  return(value)
}

#CREAZIONE R QUADRO DI NAGELKERKE: Rn
Rn = function(fit,distr){
  fam = distr
  dati = fit$model
  n = fit$df.null+1
  fit0 = update(fit,.~1, family=fam, data=dati)
  value = (1 - exp((2/n)*(logLik(fit0)-logLik(fit))))/
  (1-exp((2/n)*(logLik(fit0))))
  return(as.numeric(value))
}

#CREAZIONE R QUADRO DI KULLBACK-LEIBLER: Rkl
Rkl = function(fit){
  {
    sse1 = fit$deviance
    sse0 = fit$null.deviance
    rsq = 1-(sse1/sse0)
    return(rsq)
  }
```

```
}

#Rkl AGGIUSTATO
Rkladj = function(fit){
  sse1 = fit$deviance
  sse0 = fit$null.deviance
  rsq.adj = 1-((sse1/sse0)*(fit$df.null/fit$df.residual))
  return(rsq.adj)
}

#CREAZIONE R QUADRO DI DABAO ZHANG: Rv

#funzione per calcolare la varianza V'(mu)
#utile per il calcolo della distanza
deriv.fun = function(x,fam){
  if(fam=="normal"){
    v2 = 0
    v1 = 0
    v0 = 1
  }
  else if(fam=="binomial"){
    v2 = -1
    v1 = 1
  }
  else if(fam=="poisson"){
    v2 = 0
    v1 = 1
  }
  else if(fam=="Gamma"){
    v1 = 0
    v2 = 1
  }
  else if(fam=="quasibinomial"){
    v2 = -1
    v1 = 1
  }
  else if(fam=="quasipoisson"){
    v2 = 0
    v1 = 1
  }
  else if(fam=="negative.binomial"){
    v2 = 1/theta
    v1 = 1
  }
  ris = (2*v2*x)+v1
  return(ris)
}
```

```
#calcolo distanza fra due punti
dist.v = function(da,db,fam){
  if(fam=="normal"){
    v1=0
    dvab = (1+v1^2)*((db-da)^2)
  }
  else if(fam=="binomial"){
    v2=-1
    dvab = (1/(16*(v2^2)))*(log((db+sqrt(1+db^2))/
    (da+sqrt(1+da^2)))+(db*sqrt(1+db^2))-(da*sqrt(1+da^2)))^2
  }
  else if(fam=="poisson"){
    v1=1
    dvab = (1+v1^2)*((db-da)^2)
  }
  else if(fam=="Gamma"){
    v2=1
    dvab = (1/(16*(v2^2)))*(log((db+sqrt(1+db^2))/
    (da+sqrt(1+da^2)))+(db*sqrt(1+db^2))-(da*sqrt(1+da^2)))^2
  }
  else if(fam=="quasibinomial"){
    v2 = -1
    v1 = 1
    dvab = (1/(16*(v2^2)))*(log((db+sqrt(1+db^2))/
    (da+sqrt(1+da^2)))+(db*sqrt(1+db^2))-(da*sqrt(1+da^2)))^2
  }
  else if(fam=="quasipoisson"){
    v2 = 0
    v1 = 1
    dvab = (1+v1^2)*((db-da)^2)
  }
  else if(fam=="negative.binomial"){
    v2 = 1/theta
    dvab = (1/(16*(v2^2)))*(log((db+sqrt(1+db^2))/
    (da+sqrt(1+da^2)))+(db*sqrt(1+db^2))-(da*sqrt(1+da^2)))^2
  }
  return(dvab)
}

Rv = function(y,fit,fam){
  fit.v = fitted(fit)
  fit0 = fitted(update(fit,.~1))
  if(fam=="normal"|fam=="poisson"){
    num = dist.v(y,fit.v,fam)
    den = dist.v(y,fit0,fam)
  }
```

```
  else{
    da = deriv.fun(y,fam)
    db = deriv.fun(fit.v,fam)
    num = dist.v(da,db,fam)
    dc = deriv.fun(fit0,fam)
    den = dist.v(da,dc,fam)
  }
  value = 1 - (sum(num)/sum(den))
  return(value)
}

#Rv AGGIUSTATO

Rvadj = function(y,fit,fam){
  n = length(y)
  p = length(coef(fit))
  fit.v = fitted(fit)
  fit0 = fitted(glm(y~1,family=fam))
  if(fam=="normal"|fam=="poisson"){
    num = dist.v(y,fit.v,fam)
    den = dist.v(y,fit0,fam)
  }
  else{
    da = deriv.fun(y,fam)
    db = deriv.fun(fit.v,fam)
    num = dist.v(da,db,fam)
    dc = deriv.fun(fit0,fam)
    den = dist.v(da,dc,fam)
  }
  value = 1 - ((sum(num)/(n-p))/(sum(den)/(n-1)))
  return(value)
}
```

Listing A.2: Simulazione data-set per calcolo dei coefficienti non aggiustati

```
Rsgen = function(Nsim=100,beta.max=8,fam="binomial",
Seed=281196,display=TRUE,compare=FALSE){
  set.seed(Seed)
  X = c(rep(1,25),rep(-1,25))
  beta = seq(0,beta.max,by=0.1)
  Rlr.final = Rn.final = Rkl.final = Rv.final =
  Rkladj.final = Rvadj.final = rep(NA,length(beta))
  for(i in 1:length(beta)){
    Rlr.gen = Rn.gen = Rkl.gen = Rv.gen =
    Rkladj.gen = Rvadj.gen = rep(NA,length(beta))
    if(i%%5==0) print(i)
```

```
  current.beta = beta[i]
  eta = current.beta*X
  if(fam=="binomial")  {p = plogis(eta)}
  else if(fam=="poisson")  {p = exp(eta)}
  else if(fam=="Gamma")  {p = 0.01/(current.beta+1+eta)}
  Y = matrix(NA,50,Nsim)
  for(j in 1:Nsim){
    if(fam=="binomial")  {Y[,j] = rbinom(50,1,p)}
    else if(fam=="poisson")  {Y[,j] = rpois(50,p)}
    else if(fam=="Gamma") {Y[,j] = rgamma(50,shape=100,scale=p)}
    current.model = glm(Y[,j]~X, family=fam)
    Rlr.gen[j] = Rlr(current.model,distr=fam)
    Rn.gen[j] = Rn(current.model,distr=fam)
    Rkl.gen[j] = Rkl(current.model)
    Rv.gen[j] = Rv(Y[,j],current.model,fam)
    Rkladj.gen[j] = Rkladj(current.model)
    Rvadj.gen[j] = Rvadj(Y[,j],current.model,fam)
  }
  Rlr.final[i] = mean(Rlr.gen)
  Rn.final[i] = mean(Rn.gen)
  Rkl.final[i] = mean(Rkl.gen)
  Rv.final[i] = mean(Rv.gen)
  Rkladj.final[i] = mean(Rkladj.gen)
  Rvadj.final[i] = mean(Rvadj.gen)
}
R.square = list(Rlr = Rlr.final, Rn = Rn.final,
Rkl = Rkl.final, Rv = Rv.final, Rkladj = Rkladj.final,
Rvadj = Rvadj.final)
if(display){ #grafici degli indici calcolati
  plot(beta,Rlr.final,type="l",lwd=1,lty=1,
  ylim = c(0,1),xlab=expression(beta),
  ylab=expression("Coefficienti R "^2))
  lines(beta,Rn.final,lwd=1,lty=5)
  lines(beta,Rkl.final,lwd=2,lty=1)
  lines(beta,Rv.final,lwd=2,lty=5)
  legend((beta.max-1.2),0.2,c("RLR","RN","RKL","RV"),
  lty=c(1,5,1,5),lwd=c(1,1,2,2), bty="n")
}
if(compare){ #confronto indici aggiustati vs non aggiustati
  par(mfrow=c(1,2))
  plot(beta,Rkl.final,type="l",lwd=1,ylim = c(0,1),
  xlab=expression(beta), ylab="Coefficiente R quadro")
  lines(beta,Rkladj.final,lwd=1,col=2)
  plot(beta,Rv.final,type="l",lwd=1,ylim = c(0,1))
  lines(beta,Rvadj.final,lwd=1,col=2)
}
par(mfrow=c(1,1))
```

```
  return(R.square)
}

#esempio
Rfinal.bin = Rsgen(beta.max=6,fam="binomial",compare=F)
```

LISTING A.3: Simulazione data-set per calcolo dei coefficienti aggiustati

```
library(rsq) #libreria per estrazione delle x

Radjtest3 = function(Nsim=100,beta.max=5,Seed=281196,fam="binomial"){
  set.seed(Seed)
  beta = seq(0,beta.max,by=0.1)
  Rkladj.final = Rvadj.final = Rkl.final = Rv.final = rep(NA,length(beta))
  for(i in 1:length(beta)){
    Rkladj.gen = Rvadj.gen = Rkl.gen = Rv.gen = rep(NA,length(beta))
    if(i%%5==0) print(i)
    current.beta = beta[i]
    if(fam=="binomial") {X = simglm(family="binomial",
    lambda=current.beta,n=50,p=3)}
    else if(fam=="poisson") {X = simglm(family="poisson",
    lambda=current.beta,n=50,p=3)}
    else if(fam=="Gamma") {X = simglm(family="Gamma",
    lambda=current.beta,n=50,p=3)}
    X1 = X$yx$x.1
    X2 = X$yx$x.2
    X3 = X$yx$x.3
    X123 = cbind(X1,X2,X3)
    eta = X123%*%as.matrix(X$beta[-1])
    if(fam=="binomial") {p = plogis(eta)}
    else if(fam=="poisson") {p = exp(eta)}
    else if(fam=="Gamma") {p = 0.01/(current.beta+1+eta)}
    Y = matrix(NA,50,Nsim)
    for(j in 1:Nsim){
      if(fam=="binomial") {Y[,j] = rbinom(50,1,p)}
      else if(fam=="poisson") {Y[,j] = rpois(50,p)}
      else if(fam=="Gamma") {Y[,j] = rgamma(50,shape=100,scale=p)}
      complex.model = glm(Y[,j]~X123, family=fam)
      simple.model = glm(Y[,j]~X123[,1], family=fam)
      Rkladjt = Rkladj(complex.model)
      Rvadjt = Rvadj(Y[,j],complex.model,fam)
      Rklt = Rkl(complex.model)
      Rvt = Rv(Y[,j],complex.model,fam)
      Rkladj.gen[j] = Rkladjt - Rkladj(simple.model)
      Rvadj.gen[j] = Rvadjt - Rvadj(Y[,j],simple.model,fam)
      Rkl.gen[j] = Rklt - Rkl(simple.model)
      Rv.gen[j] = Rvt - Rv(Y[,j],simple.model,fam)
```

```
    }
    Rkladj.final[i] = mean(na.omit(Rkladj.gen))
    Rvadj.final[i] = mean(Rvadj.gen)
    Rkl.final[i] = mean(Rkl.gen)
    Rv.final[i] = mean(Rv.gen)
  }
  R.square = list(Rkladj = Rkladj.final, Rvadj =
  Rvadj.final, Rkl = Rkl.final, Rv = Rv.final)
  return(R.square)
}
```

Appendice B

Applicazione su data-set reali

Listing B.1: Applicazione coefficienti di determinazione su data-set reali

```
#esempio modello di quasi-verosimiglianza

rats = read.table("Rats.dat",header=T)
attach(rats)

y = s/n

rats0 = glm(y~group*h, weights=n, family=binomial)

pers.res = resid(rats0, type="pearson")
phitilde = sum(pers.res^2)/rats0$df.residual
# phitilde = 2.75 >> 1

#STEP 1
rats0.ql = glm(y~group, family=quasibinomial, weights=n)
summary(rats0.ql)
Rv(y,rats0.ql,fam="quasibinomial")
Rvadj(y,rats0.ql,fam="quasibinomial")

#STEP 2
rats1.ql = glm(y~h, family=quasibinomial, weights=n)
summary(rats1.ql)
Rv(y,rats1.ql,fam="quasibinomial")
Rvadj(y,rats1.ql,fam="quasibinomial")

#STEP 3
rats2.ql = glm(y~group+h, family=quasibinomial, weights=n)
summary(rats2.ql)
Rv(y,rats2.ql,fam="quasibinomial")
Rvadj(y,rats2.ql,fam="quasibinomial")
```

```
#STEP 4
rats3.ql = glm(y~group*h, family=quasibinomial, weights=n)
summary(rats3.ql)
Rv(y,rats3.ql,fam="quasibinomial")
Rvadj(y,rats3.ql,fam="quasibinomial")
```

Bibliografia

A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, 1996.

P. Allison. What's the best R-squared for logistic regression?, 2013. URL `https://statisticalhorizons.com/r2logistic`.

N.R. Draper e H. Smith. *Applied Regression Analysis, 3rd Edition*. Wiley, 1998.

S. Kullback R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

G.S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University, 1983.

L. Magee. R^2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician*, 44:250–253, 1990.

C.N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10:65–80, 1982.

N.J.D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991.

K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

L. Pace A. Salvan. *Introduzione alla statsitica. II Inferenza, verosimiglianza, modelli*. CEDAM, 2001.

P.K. Dunn G.K. Smyth. *Generalized Linear Models With Examples in R*. Springer, 2018.

D.R. Cox E.J. Snell. *The Analysis of Binary Data (2nd ed.)*. London:Chapman and Hall, 1989.

L. Fahrmeir G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 2001.

M. Grigoletto F. Pauli L. Ventura. *Modello lineare. Teoria e Applicazioni con R*. Giappichelli, 2017.

A.C. Cameron A.G. Windmeijer. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77:329–342, 1997.

D. Zhang. A coefficient of determination for generalized linear models. *The American Statistician*, 71:310–316, 2017.